DOCTOR OF ENGINEERING (ENGD)

Compensating for distance compression in virtual audiovisual environments

Finnegan, Daniel

Award date:
2017

Awarding institution:
University of Bath

Compensating for Distance Compression in Virtual Audiovisual Environments

Daniel Joseph Finnegan

A thesis submitted for the degree of Doctor of Engineering

University of Bath

Centre for Digital Entertainment

April 2017

Abstract

Virtual environments are increasingly being used for various applications. In recent times, with the advent of consumer grade systems, virtual reality has reached a critical mass and has exploded in terms of application domains. Extending from games and entertainment, VR is also applied in military training, remote surgery, flight simulation, co-operative work, and education. While all of these applications require careful design with respect to the interaction and aesthetics of the environment, they differ in their requirement of veridical realism: the impression of suspending disbelief to the point where perception in the environment is equal to the real world. At the same time, research in human centred disciplines have shown predictable biases and 'errors' in perception with respect to the environment intended by the designer. This can be a challenge when certain perceptual phenomena prohibit the applicability of VR due to a discontinuation in what is rendered and what is actually perceived by the observer.

This thesis is focused on a specific perceptual phenomenon in VR, namely that of distance compression, a term describing the widespread underestimation of distances that occur in VR relative to the real world. This perceptual anomaly occurs not only in visual based virtual environments, as compression has been observed and studied in auditory only and audiovisual spaces too. The contribution of this thesis is a novel technique for reducing compression, and its effectiveness is demonstrated in a series of empirical evaluations. First, research questions are synthesized from existing literature and the problem is introduced and explained through rigorous review of previous literature in the context of spatial audio, virtual reality technology, psychophysics, and multi-sensory integration. Second, the technique for reducing distance compression is proposed from an extensive literature review. Third, the technique is empirically tested through a series of studies involving human participants, virtual reality hardware, and bespoke software engineered for each study. Finally, the results from the studies are discussed and concluded with respect to the research questions proposed.

Contents

List of Figures

List of Tables

Publications

The work in Chapter 5 led to two publications:

Daniel J. Finnegan, Eamonn O'Neill, and Michael J. Proulx. Compensating for Distance Compression in Audiovisual Virtual Environments Using Incongruence. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA.

Daniel J. Finnegan, Eamonn O'Neill, and Michael J. Proulx. An Approach to Reducing Distance Compression in Audiovisual Virtual Environments. *In Proceedings of the 2017 IEEE* 3^{rd} *VR Workshop on Sonic Interactions for Virtual Environments (SIVE).* IEEE, New York, NY, USA.

The genesis of software used in Chapter 6 came from the following publication:

Daniel J. Finnegan, Eduardo Velloso, Robb Mitchell, Florian Mueller, and Rich Byrne. Reindeer & wolves: exploring sensory deprivation in multiplayer digital bodily play. *In Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play (CHI PLAY '14).* ACM, New York, NY, USA.

Preface

I'll begin with an anecdote. Towards the end of my EngD, I remember having a discussion with my academic advisor, Prof. Eamonn O'Neill regarding a recent technology demo he'd played with. The system in question was the HTC Vive, a new [1] virtual reality (VR) system co-developed by Valve, of Half-Life© fame. Eamonn said he was playing around with the headset, and tried reaching for an object in the environment. After many failed attempts at grasping the object, he realised he was undershooting his aim. Every time he would reach, he would not extend his arm long enough, and therefore the tracking system that mapped his physical arm to a virtual arm was not moving the virtual representation of his arm in the virtual world far enough. He described it as though he was *compressing* the distance to the target object in the virtual world. We both laughed for a moment, but I'll never forget the moment I felt my EngD vindicated.

This thesis regards a well studied problem in VR systems; namely the issue of distance perception being skewed towards the observer, a compression of the distance between the observer and some object. While a problem worthy of consideration due to practicalities of VR application, I also think this is an interesting problem from a purely curious perspective. Why, after years of engineering breakthroughs in headset design, graphical display hardware, and software rendering does this problem still occur? What, if any, are the perceptual factors causing this problem? Finally, what can we do about it?

That final bit is important and it is what I set out to do for the past 4 years. At first I thought perhaps this was a hardware problem and that the issue was with

[1] At the time of writing; depending on when you're reading this it may be old news or retain legendary status

the fact that researchers were using different–i.e. older–equipment than the state-of-the-art we have today. This hypothesis was debunked relatively painlessly by reading half a dozen papers that spanned well over 10 years. Researchers studying distance perception with modern hardware systems saw the same compression phenomenon as their predecessors. Next I thought perhaps it was an issue with the way head mounted displays (HMDs) operate, how they split light across a collimated display which would cause issues with how our eyes naturally converge to a point in the real world. This too was debunked by the observation of the compression phenomenon in non HMD systems such as CAVEs (Audiovisual Experience Automatic Virtual Environment) or LSIDs (Large Screen Immersive Displays).

After doing some more digging, I began to consider more in-depth theories that led me into studying the problem from the perspective of a psychologist trying to understand human perception. Concurrently to this, there was the observation of distance compression in virtual auditory environments, not just traditional visual VR. I decided to follow this path, and the result was an exploitation of multi-sensory perception. This exploitation led to my contribution of compensating for distance compression in audiovisual virtual environments through cross-modal incogruence.

An EngD is a personal journey, driven by curiosity, fuelled by passion, and supported by people. First and foremost, I thank my academic advisors Prof. Eamonn O'Neill and Dr. Michael Proulx. In the US the term 'advisor' rather than 'supervisor' is used. I feel this is a more appropriate term as postgraduate research should be largely autonomous. Rather than supervision, I always sought advice, which Eamonn and Michael both gave endlessly. To Eamonn, I am indebted for all you have taught me. Not just with regards to HCI research and related work, but also general research conduct, critical review & analysis, and a healthy dose of scepticism Your wisdom extends far beyond the field of HCI and together with Michael you have shaped me into the critical thinker I am today. To Michael, your constant positivity helped keep me afloat at times when I felt overwhelmed. Your vast knowledge and ability to pull an appropriate citation out of the blue during all of our group conversations has given me insight and enabled me to expand the domain of my work. You stimulated my interest in

psychology, always fascinating me with insights from the field that broadened my thinking and enabled me to conduct better research in a truly cross-disciplinary environment. I am lucky to have had you both as my advisers, and our relationship has developed so that I see you both not just as advisers, but also as friends.

I extend my gratitude to Nicky Birch, Neville Daniel, Nigel Brown, and Antoine Pastor; each of whom I worked in direct contact with during my time at Somethin' Else and who all enriched my studies with their industrial expertise. To Rob McHardy, my industrial advisor, for making me a better software engineer. I have learned innumerate lessons from you all and I carry them forward into my research and development career. I thank my mother and father Elaine and Mel, my brothers Benjamin and Joshua, and my sister Rebecca. All of you have been supportive in my studies, encouraging me to continue and constantly asking how things were going. It took a long time and a lot of work, but I made it in the end.

Chapter 1

Introduction

"I think you'll find its a bit more complicated than that"

Ben Goldacre

VR (Virtual Reality) has emerged as a major paradigm in the gaming industry in recent years. With the advent of consumer grade, cheap and mass produced head mounted display technology, VR has become more accessible [3]. This in turn has driven demand for applications, but the advent of developer tools that ease the work flow in VR content production has enabled developers to somewhat meet such demands. While VR is not new, it is certainly new to the consumer market. However, VR has also attracted attention from other consumer contexts outside of the games industry. The film industry, aircraft design and engineering, medical, and even education industries can all benefit from the application of head mounted displays (HMDs).

Before discussing the various topics in this thesis, it is necessary to first introduce some definitions. Burdea and Coiffet describe VR in terms of functionality, as a "simulation in which computer graphics is [sic] used to create a realistic-looking world" [4]. This is a very naïve definition however, as while the claim that a reality can be simulated through computer graphics and visual stimulation, the result would be a very poor simulation of the reality we exist in. My own definition of VR is as follows: VR is a spectrum of simulations of sensory experiences. Any

simulated experience that captivates at least one human sensory modality lies on this spectrum, and it is populated with simulations that capture not only vision, but audition, touch, smell, and taste in various combinations. At the near end of the spectrum are uni-sensory simulations: experiences where just a single sensory modality is accommodated. At the far end is any simulation that fully accommodates all human sensory modalities. The goal of VR research and development is to reach the far end of this spectrum.

Traditionally VR was in line with Burdea and Coiffet's definition, driven by advances in graphics technology. While we are still a long time from reaching the end of the spectrum, in recent years VR has shifted in to simulations of multi-sensory experiences. Sound, smell, and taste have all been incorporated with varying degrees of success (see Chapter 3, Section 3.2.1). While vision remains the dominant domain of research in VR, recently audition has entered the race as sound rendering technology and techniques have matured greatly. Of these, spatial audio techniques and technology have risen to the forefront of VR research in tandem with graphics. Spatial audio is audio that is rendered in 3D. Just as graphics technology give a third dimension to what's seen on screen, spatial audio enables sound to come from various directions around the listener, just as it does in the real world.

Another definition necessary before our discussion is that of perception: the experience of sensory information that results in an understanding of a given environment [5]. VR is the result of technology and perception working together to drive any simulated experience. In order to reach the goal of VR research above, psychological research on perception needs to continue in parallel. Human-computer interaction (HCI) aims to incorporate the human agent into the design of a system; indeed the human is a core component of the interactive system. From this human centred design perspective, it is imperative to understand human reasoning and decision making, interaction paradigms such as hands free, eyes free, and computer supported co-operation along with systems engineering and software implementation. However, if VR is to reach its full potential and we wish to perform work while *immersed* in a virtual environment, then it is imperative that we also study perception.

Of the many perceptual processes, one of utmost importance is that of spatial perception. Extending past the notion of where objects are situated relative to one another in a 2D setting, consider the notion of distance perception. Over the years, a consistent finding in the literature is that egocentric distance perception follows a power law with respect to increasing distance. Interestingly, this perception seems to be independent of sensory modality, operating in both audio and visual environments. It has been studied substantially in the HCI and Psychology literature, yet a complete model detailing its cause is yet to be derived. Many studies have simply observed it across various contexts [6, 7, 8, 9]. Due to the large number of factors involved (See the work from Renner et al. for a review [2], and Chapter 4), it is very difficult to describe a fully complete model.

Take for example the use of VR to facilitate remote surgery. While wearing a headset, the surgeon sees a real life model of the patient, and through the application of teleoperation, control a pair robotic hands to perform the removal of a tumour. The VR environment enables a heads-up display (HUD) for meta information such as the patient's vitals, medical history etc. As the surgeon operates, her hand-eye coordination is put to the test. However, this is only valid providing that the VR environment has not led to a *mis-mapping* of her hand-eye coordination. How deep she would normally cut depends on how she has adapted over many years of practice to make a cut based on her depth perception. Given that this is skewed in VR, might the surgeon cut too close, or overcompensate and cut too far?

The problem of distance compression has been tackled by previous researchers through manipulation of the auditory and visual display [10, 11]. These manipulations have been applied in isolation of one another. As VR applications become more and more progressively multi-sensory, these uni-modal solutions will not suffice. We perceive our environment from multiple sensory streams, and make judgements by iterating over all information and integrating all together to arrive at an an informed decision. This thesis explores manipulation of an audiovisual environment.

1.1 Research Goals & Methods

My research goals are to understand distance compression in virtual environments. The problem is multi-faceted, with numerous factors described over the years by various researchers (See Chapter 4, Section 4.2.1). My contribution in the form of a novel correction technique, independent of hardware employed and software algorithms, is simple to implement. It is applied to different virtual scenarios, demonstrating flexibility and versatility. I verify it initially through rigorous psychophysical methods in a simplistic, abstract environment, before scaling up to environments with more distance cues and enabling the observer to interact in more natural ways.

1.1.1 Research Questions

RQ1: Can distance compression be compensated for in audiovisual virtual environments using incongruence?

My approach to answering this first question draws upon previous research that observed human perception of distance in virtual environments. By studying the response of participants in controlled environments, the perceptual distance of a given stimulus was recorded and fit to a model that explains the perceived distance with respect to minimal distance cues. Rather than trying to identify and fully explain the underlying mechanism, my novel approach takes this model and inverts it: in other words, we compute an *actual* distance value for a given *perceived* distance value.

RQ2: Does the compensation function generalize to less abstract environments?

After manipulation and observing an effect, it is important to reflect on the experimental design and scenario used. In the first study a simple, abstract environment was used. However, such a simple environment is unlikely to be used in a real application. Thus, the second question concerns ecological validity: how well

does the distance compression function operate in less rigid (i.e. more realistic) environments? This question was handled by designing an experiment that incorporated photo-realistic visual cues to distance and modified the experimental task from the first experiment.

RQ3: Does the distance compression compensation function generalize to dynamic environments?

While RQ2 addresses ecological validity with respect to external cues from the environment, VR has evolved away from being a passive medium. Interaction in VR more than merely tracking head rotations is critical to an immersive experience, and in the consumer market at least, there already exist systems which provide full body motion interaction while wearing a headset[1]. Another question to pose then is whether the distance compression compensation function can also reduce compression in environments which enable full body motion. To address this, I developed a small motion tracking system enabling participants to physically move in the real world, with their movements being mapped into the virtual world.

1.2 Thesis Overview

This thesis is split into 2 main parts. To begin, Part I gives a background review on the basics of sound signal processing, introducing core concepts of spatial audio including Fourier transforms, convolution, and binaural rendering. Then I move on to discuss psychophysics, which discusses how science regards the human brain as a 'black box'. This description lends itself nicely to the field of signal processing and reverse engineering; you observe a system's output for a given input, then try to manipulate the input in a controlled manner in order to observe a predicted output. If your predictions are correct, you can learn a model of the system. This model can then be tested by formulating hypotheses, then performing some manipulation, much like standard Null Hypothesis Significance Testing (NHST). Psychophysics borrows techniques like Signal Detection Theory

[1]For example: `http://www.virtuix.com/`

from Signal Processing in order to understand how the brain processes low level stimuli streams through multiple sensory inputs (vision, audition, smell etc.) and develop theories of perception based on these studies.

In Part II I introduce the literature on distance perception in virtual environments. I discuss the various audio and visual cues that are known to contribute to distance perception, before discussing the phenomenon of distance compression. With respect to previous studies, I give a background into distance perception, and note some techniques that have been applied in the past in order to manipulate the compression effect in order to prevent, or at least reduce it. I then detail the studies that I have designed myself that make up the core of my contribution. Each study is based around the concept of reducing distance compression through a novel compensation technique, and builds on top of the previous study. The technique is grounded in a well supported theory of multi-sensory integration, and involves rendering the visual and the audio displays incongruently to one another. This incongruence produces a counter bias, affecting the audiovisual perception of distance towards a target in a virtual environment. The net result is a reduction in the compression error observed in a virtual environment. I describe the custom software that I implemented in order to conduct my experimental designs in brief during the main chapters, leaving core implementation details to an appendix for interested readers.

1.2.1 Chapter Breakdown

In order to convey the idea of cross-modal compensation, it is necessary to provide some primers to existing work in binaural audio, digital signal processing, perception, psychophysics, multi-sensory integration, and finally distance compression itself. Thus, I have broken my thesis into chapters that present the necessary background work as a pre-requisite for appreciating my own work:

1. Chapter 2 presents the basics of signal processing, before moving to the convolution operation. Convolution involves the frequency combination of two signals that forms the basis of binaural spatialisation over headphones, a technology that I used to base my incongruence work on. Chapter 2 then

begins to detail perceptual studies relating to localization of virtual auditory and visual sources, and critiques some of the studies in order to define some of the factors associated with localization and distance perception in both audition and vision.

2. Chapter 3 introduces psychophysics and details signal detection theory, before introducing the topic of multi-sensory integration in the context of audiovisual stimuli mainly, but includes relevant studies in audio-haptic and visuo-haptic perception in order to give a grander picture of multi-sensory perception.

3. Chapter 4 provides a literature review of past studies which have observed the phenomenon of compression with respect to distance perception. In it, I provide a broader context for the view, and slowly funnel towards the idea of correcting the compression issue.

4. Chapter 5 & Chapter 6 detail studies I conducted which applied all the theory and knowledge discussed in the previous chapters to distance compression in audiovisual VR, especially by applying the techniques described in Chapters 2 & 3.

While each chapter is self-contained where possible, some sections will reference previous or future chapters for further reading. Thus, the thesis is recommended to be read in order rather than individual chapters isolated from the rest. The final chapter concludes by bringing together all the core concepts from previous chapters and specifying future work.

Part I

Signal Processing & Psychophysics

Chapter 2

Spatial Audio & Digital Signal Processing

"If we judge ourselves by what is hardest for us, we may take for granted those things that we do easily and routinely"

Nate Silver

2.1 Introduction

Digital Signal Processing (DSP) is an integral part of 3D audio, as it details methods and techniques for processing audio that allow the transformation from a 2D monaural signal to a 3D binaural signal over headphones, or a multi-speaker rendering system for creating virtual audio images in a real space. It helps to have an understanding of the fundamental concepts in audio signal processing and auditory signal perception before moving on to discussing auditory distance perception, as many of the aspects of auditory distance perception derive from factors in the processing and characteristics of audio signals. In order to achieve this, I have structured this chapter into 3 main parts.

First, I discuss the basics of an auditory signal, from its waveform representation in a computer, to intensity and frequency components of a waveform. I then move

onto some perceptual aspects of audio signals, such as pitch, timbre, and how these aspects have been exploited at the user interface. From here, I discuss some higher level perceptual ideas that are based on the fact that we hear binaurally; topics such as localization based on inter-aural time differences and inter-aural level differences, auditory scene analysis which focuses on discriminating between multiple sound sources and interpreting an acoustic scene. I then introduce some rendering methods for spatial audio. First, I begin with loudspeaker methods; stereo and surround sound as the basic channel delivery techniques for spatial sound. Next, I move on to more advanced techniques such as Ambisonics, Vector Based Amplitude Panning (VBAP), and Wave Field Synthesis (WFS). After this, I then return briefly to DSP in order to discuss a fundamental operation called convolution, which is involved in the spatialization of audio over headphones for virtual 3D audio. Convolution with the Head-Related-Transfer Function (HRTF) forms the process of binaural audio, a form of 3D audio used in the studies I report in this thesis. I compare different flavours of binaural audio: personal HRTF models to generic and then even synthesized ones.

Finally, I discuss spatial audio and its role in distance perception studies. Various models for distance perception have been proposed, based on the environment of the listener. By the end of this chapter, the reader will have the background necessary to interpret the content of Chapters 4, 5, & 6.

2.2 Sound Perception: From Pressure to Information

2.2.1 The Waveform

An audio signal is represented as at least a 1 dimensional vector of numbers representing volume levels as samples. These samples are played back through an electrical transducer that turns an electrical impulse into a vibration via a motor, which produces pressure waves that we then perceive as sound. Figure 2-1 represents a simple sinusoidal waveform. Samples are played at a designated sample rate, which determines how many samples are played a second. The waveform of a signal can be described by two main attributes; the amplitude

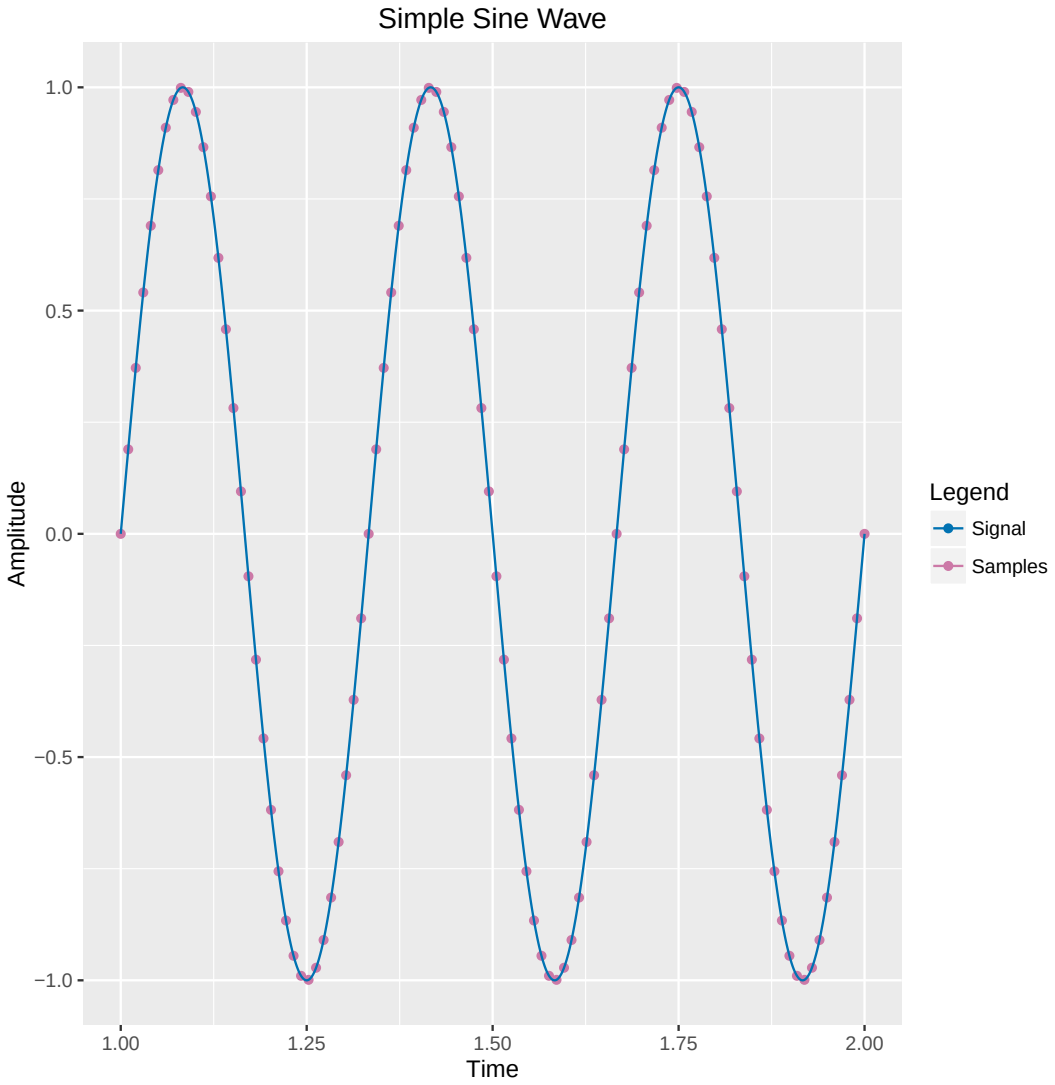

Figure 2-1: A typical waveform plot showing the intensity, frequency, and phase properties. For details on source code that produced the graph, See Appendix B.

or the change in pressure through time, and the frequency, which defines the repetitive nature of the signal. This change in pressure transmits through the ear canal where the malleus, incus, and stapes (more commonly referred to as the hammer, anvil, and stirrup) transduce the waveform to an electrical signal that is then passed through the cochlear nerve into the brain, where it is processed by the human auditory system (HAS). The frequency is the pattern of the amplitude change. Measured in hertz (Hz), all periodic waves repeat their pressure shift pattern at the rate specified by the frequency.

Another attribute of the waveform is the phase, which simply describes the difference between the waves starting amplitude and a known reference point in time. While phase plays more of a role in digital signal processing, it has important repercussions for auditory localization studies. As phase applies to periodic waveforms, for low frequency signals it can be a cue for localization, as the time difference is more noticeable [12, 13]. This plays an essential role in human perception of sound, which I'll discuss in more detail in Section 2.5.2.

2.2.2 Loudness, Pitch, and Timbre

While amplitude, frequency, and phase are physical characteristics of a sound, when discussing perception, it is more common to talk in terms of loudness and pitch. Another term often used is timbre which relates to the quality of a sound, making it identifiable in relation to different instruments. Loudness is the perceptual counterpart to amplitude. While amplitude is a linear scale, loudness on the other hand is not. The decibel system, which follows a logarithmic scale, was devised to account for the high dynamic range of the HAS. It represents the ratio between the current pressure and a reference point of pressure in air, as shown in Equation 2.1.

$$L = 20 \log_{10} \left(\frac{p_c}{p_{ref}} \right) \ dB \tag{2.1}$$

where p_c is the pressure of the current wave. This equation maps amplitude levels non-linearly to loudness levels, due to the nature of human perception of amplitude.

Pitch is the psychoacoustic term for the perception of a sound's frequency. As it is subjective, it is difficult to quantify or rather measure objectively. The ability to *discriminate* pitches is measurable however, along with the ability to rank and categorize them as low or high pitched. Some people are believed to have the ability of perfect pitch, where they can ascertain the true pitch of a sound in the absence of a reference tone for comparison [14].

Finally, timbre can be defined as the attribute of a sound where, in comparison to another sound of the same pitch and loudness, is perceived as different [15]. A guitar playing a C note sounds different to a flute playing the same note;

their timbre are different. In HCI timbre, is most useful applied as an interface mechanism in earcons: melodic samples that are used to hierarchically represent information [16]. McGookin details the application of earcons to represent data, and discusses the limits of their applicability [17]. One thing he notes is the striking inability of humans to distinguish harmonics from the fundamental frequency. When people hear a mixture of harmonic notes, they believe they have heard the base frequency, even when it is not present. The example of a note made up from 400Hz, 1200Hz, and 1400Hz (all harmonics of a 200Hz tone), to the listener's ears will sound of the 200Hz tone.

Given such perceptual counterparts to characteristics of an auditory signal, we may begin to hypothesize how these characteristics interact with one another. In doing so, we discover many peculiar and interesting effects of loudness on pitch, loudness discrimination, and even multi-modal interactions such as bright light and high pitched sounds (as discussed above). In the following sections I give a little more detail into these perceptual artefacts, and discuss the consequences they have on designing effective auditory interfaces and interactive scenarios in VR.

2.2.3 Loudness Constancy

The perceptual properties of loudness vary significantly from the physical properties of amplitude, due to the logarithmic nature of the HAS's sensitivity to loudness. Loudness perception is well fit to a power law of the form given in Equation 2.2.

$$L = kI^{\alpha} \tag{2.2}$$

Zahorik shows how the k and α values are typically derived from empirical psychoacoustics data, with α generally taking the form of 0.3 [7]. As loudness is the perceptual sister to amplitude, the expected result of a drop or increase in the amplitude of a signal would be a similar drop or increase in the perceived loudness of the signal. However, an interesting perceptual phenomenon known as loudness constancy is observed, where varying changes in the power output of a sound source results in a constant perceived loudness at the ears. Zahorik

details two hypotheses regarding loudness constancy. The first suggests an influence of memory on the perception of loudness. As the ears become familiar with the power output of a sound source (e.g. the intensity of a jet engine, or the engine of a car, or indeed the sound of a familiar voice), the sensitivity to intensity reduces, forming a constant perceived loudness [7, 9].

The second hypothesis is based on the concept of auditory localization. If the brain can determine the distance towards an object of unknown power, then this distance estimation may be passed as input to the loud constancy function. This would imply that the brain can modulate the perceived intensity of a sound source at a distance so that it is perceived as constant. Barron performed a series of analyses on perceived constancy of loudness in concert halls, and found that the regression model for loudness was improved when distance was accounted for [18]. He notes an interesting observation in his findings: namely that the direction of the coefficient for distance implies that loudness was perceived greater as distance was increased. Linking with the second hypothesis from Zahorik, it is believed that in concert halls, listeners' perceived loudness depends on how distant they perceive themselves to be from the stage [18]. This perceptual ability is not quite clear, yet it implies some sort of visual influence on the perceived distance. Vision has been shown to impact perceived auditory distance in trials involving dummy loudspeakers [19]. This poses an interesting question: whether or not a semantic mapping is required between the visual sound source representation and the sound stimulus itself (i.e. the sound of a loudspeaker versus an abstract object that emits a sound) in order to determine distance? I'll discuss findings relating distance perception with cross-modal influences in Chapters 4, 5, & 6.

2.2.4 Interactions between Pitch and Loudness

Pitch is known to impact perceived loudness in the human auditory system. Melara & Marks for example sought to identify the interactions between pitch and loudness in experiments where participants were tasked to classify the changes in both dimensions via a timed classification (speeded) task [20]. They operationalize their study by categorizing three distinct presentation styles in terms of their impact on perceptual primacy. They discuss how dimensions may be processed ei-

ther as separable or integrable: separable when both dimensions can be streamed in separate, non-interfering channels, and integrable when some crosstalk occurs across channels. Such an interference is known as Garner Interference [21]. Thus, the three presentation styles are:

1. Baseline: One dimension is manipulated, while the other is held at a constant value.

2. Filtering: One dimension is manipulated while the other changes in an incongruent fashion (e.g in the opposite direction, at a slower/faster rate).

3. Correlate: Both dimensions change at the same rate, and in the same direction.

Melara & Marks found significant interaction effects between timbre and loudness where both dimensions appeared to support concurrent processing when presented congruently and when correlated. However, the effect disappeared when the two interfered when uncorrelated. For timbre and pitch, participant's performance was the same regardless of the correlation between both dimensions; namely it was improved with respect to a faster response time in the classification task.

Neuhoff et al. studied interaction between pitch and loudness in the context of sonification of data [22]. In their first experiment, they presented participants with a set of auditory stimuli that changed in intensity and/or frequency, in the form of a triangle wave pattern. The experiment was a divided attention task; participants had to pay attention to concurrent changes in both dimensions of pitch and loudness. This set was then presented in a manner so that participants heard both congruent dimensional change (i.e. increasing pitch & increasing loudness; decreasing pitch & decreasing loudness) and incongruent dimensional change (i.e. increasing pitch & decreasing loudness; decreasing pitch & increasing loudness). The dependent variable of *global* change to the sound was recorded, using a proxy of an analogue scale to map the change. Their results show that participants were more sensitive to concurrent increases in pitch and loudness than decreases, and that rising loudness with falling frequency was perceived as a bigger change than falling intensity and rising frequency [22]. Such an

asymmetrical finding has been found in other studies, most notable those that looked at looming scenarios: scenarios where objects move towards an observer rather than away [23, 9].

Given that a high intensity sound can be thought of as close-by, and that softer, quieter sounds may be perceived as far away, other researchers looked at the effect of emotive affect on loudness perception. Gagnon and colleagues found that fear greatly reduced the perceived distance to an object but only when that object was represented as an auditory stimulus[1]. The effect did not carry for objects represented visually. Given these emotive factors in the looming problem, it is difficult to interpret some results that do not explicitly control for other known factors in loudness-pitch interactions, as well as cross-modal interactions such as the visual and auditory looming research. While Neuhoff did not control for emotive aspects, it may be argued that participants were not specifically *induced* to a fearful state, yet it is not enough to presume that large intense sounds, i.e. the experimental stimuli themselves, did not induce this unintentionally.

2.3 Signal Processing and Spatial Rendering

As discussed so far, it should be clear to the reader that auditory perception is a well researched topic. From a signal processing point of view, we can take advantage of such interactions in order to create *digital filters*. By manipulating an audio signal with respect to its intensity and spectral content, we can directly manipulate the expected perceptual response from the listener. Namely, we can manipulate the intensity of the signal in order to impact the inter-aural level difference, and manipulate the phase and spectral content to impact the monaural cues caused by impact with the listener's outer ear (or pinna). Both ILD & monaural cues are discussed in more detail in Section 2.5. Thus, I know turn to advances in the field of digital signal processing and look at how various systems implement such digital filters in spatial (3D) audio synthesis and reproduction.

It is possible to reproduce a real world acoustic scene by processing the audio

[1]Often referred to throughout the scientific literature as *distance compression*, I discuss this phenomenon in great detail in Chapter 4.

data before rendering. This is a rather processor intensive operation, but as machines have become more powerful in recent years, it has become possible to perform these operations in real time on both desktop and mobile processors. In the following few sections, I introduce spatial audio concepts and elaborate on the technology involved in implementing spatial audio.

2.4 Spatial Audio With Loudspeakers

Spatialisation of virtual sound is a multi-faceted problem, drawing research from digital signal processing theory and human perceptual studies on sound source localization. This has led to a broad research field interest ranging from engineering, psychophysics, and human-computer interaction. On the psychophysics side, many studies have been performed in auditory localization [24, 25, 26, 27, 28, 29]. These studies have focused on understanding the perceptual response to sound sources rendered physically from a loudspeaker or virtually over headphones, using various techniques discussed in the following sections. On the engineering side of things, researchers have tackled problems around fidelity, processing latency, and channel versus object based spatialisation [12, 30, 31, 32]. The next few sections give an introduction to various audio systems, describing techniques for spatial audio rendering.

2.4.1 Stereo and Surround Sound

Given any free field space, a virtual source can be rendered at a given location by placing a transducer at that location in space (with respect to audio, this is manifested through the application of a loudspeaker connected to some processing unit for audio output). The basis of surround sound follows this approach by creating a loudspeaker 'configuration': by positioning loudspeakers in a given configuration, a soundscape can be rendered by processing the audio, mixing to form channels that are then fed to individual speakers in the configuration.

The basic form can be reduced to a single loudspeaker: the speaker can be placed in front, above, below, or to the sides of the listener in order to generate an audio

image from the respective position. Taking a step up, a stereo speaker system can be created through the introduction of a second speaker. Though seemingly unimpressive, interesting soundscapes can result from the inclusion of an extra speaker. For example, by adjusting the gain values of each speaker relative to one another, a *phantom* image can be created: this is an audio image resulting from a location somewhere in between the two physical speaker locations (See Section 2.4.2). This immediately equips the soundscape designer with 1 degree of freedom; once the location of the speakers has been fixed, the sound can then pan back and forth between both speakers. It is not difficult to imagine this technique scaling for any arbitrary amount of speakers; increasing the number of speakers in a line increases the width of the potential locations with which to position a phantom audio image.

Another example is surround sound, where main voice audio comes from the two front speakers with ambience and out-of-shot audio coming from the side and rear speakers. Low frequency sound emanates from the middle speaker in the front. This enables a sound designer to create a multichannel mix that creates a soundscape containing spatial information by mixing across the various speakers available in the configuration. One obvious drawback of this is that, if the configuration changes, then the audio must be remixed to match the new configuration. To further exacerbate things, for dynamic soundscapes, things become a little more complicated, especially considering when listeners wish to move around and interact with the sound. Stereo and Surround sound configurations are constrained by an auditory 'sweet spot': the position in the free-field space which results in the optimal audio image being generated [33].

Nevertheless, there are a few techniques using speakers that are a little more flexible than stereo and surround sound. The ones I discuss over the next few sections are Vector Based Amplitude Panning and Wave Field Synthesis.

2.4.2 Vector Based Amplitude Panning

Pulkki introduced the concept of a spatial audio production system that was agnostic to the number of speakers available [34]. His technique is based on

panning between speakers in a 3D co-ordinate system by computing vectors from the listener to the expected audio image, hence termed Vector Based Amplitude Panning (VBAP). VBAP operates over an arbitrary number of loudspeakers by modelling the sound field produced as a series of vectors between the listener and the speakers that do exist in the environment, modulating their gain values to produce phantom/virtual speakers.

Pulkki presented the panning formulation in both 2D and 3D spaces. The gains from the loudspeakers focuses on setting a power constant, called C, that satisfies Equation 2.3:

$$\sum_{1}^{n} g_n^2 = C \tag{2.3}$$

where g_n is the amplitude of speaker N. The crucial point behind the panning method is the stereophonic law of sines, or Bauer's Stereophonic Law, as shown in Equation 2.4.

$$\frac{\sin\varphi}{\sin\varphi_0} = \frac{g_L - g_R}{g_L + g_R} \tag{2.4}$$

Equation 2.4 specifies that a stereo image can be produced between a pair of speakers by taking the angle between the speakers and the x axis (φ), and the angle between the direction of the sound source and the x axis (φ_0). In 2D stereo panning, the X axis spans the listener's orientation. Providing that the loudspeakers lie on the same plane, equidistant from the listener, this is sufficient to create a stereo image emanating from an arc around the listener between two speakers [34].

Pulkki extended this panning method by formulating the gain equation as a list of vector bases. Unit vectors L_1 and L_2 from the origin (chosen to be the 'sweet spot' where the listener is positioned), with appropriate gain values applied as scalar values to the vectors. Thus, any point P from which a virtual audio image is to be rendered can be specified as a sum of these two speaker vectors (Equation 2.5).

$$P = g\bar{S}_L + g\bar{S}_R, \quad \bar{S}_L = [S_{lx}, S_{ly}]^T, \quad \bar{S}_R = [S_{Rx}, S_{Ry}]^T \tag{2.5}$$

Pulkki demonstrates how the gain value for any speaker can be computed by solving the linear equation of the virtual source point and the matrix formed

from the multiplication of the loudspeaker configuration as in Equation 2.6.

$$\bar{g} = p^T S_{LR}^{-1} = [p_x, p_y] \begin{bmatrix} S_{Lx} & S_{Rx} \\ S_{Ly} & S_{Ry} \end{bmatrix} \tag{2.6}$$

In three dimensions, the vector equations scale simply by adding a new element to the p and g vectors, producing Equation 2.7.

$$\bar{g} = [p_x, p_y, p_z] \begin{bmatrix} S_{Lx} & S_{Rx} & S_{Mx} \\ S_{Ly} & S_{Ry} & S_{My} \\ S_{Lz} & S_{Rz} & S_{Mz} \end{bmatrix} \tag{2.7}$$

In Equation 2.7, S_M represents the loudspeaker that is placed above the other loudspeakers to incorporate elevation into the phantom image. The same method for scaling the number of loudspeakers used also applies in three dimensions; rather than choosing a pair of speakers and computing the gain factors to produce an image along the inter-loudspeaker arc, a set of 3 speakers are chosen at any one time, producing a spherical triangle space within which an image can be produced.

While focused on the hardware and formulation of VBAP, it is important to consider the perceptual quality of the panning technique. Loudspeaker panning has been reviewed with respect to localization. Cofino et al. compared VBAP to a simple linear panning (LP) technique [35]. They designed an experiment where participants were asked to localize a sound in the frontal plane, in a 5AFC task (See Chapter 3, Section 3.3.1 regarding nAFC tasks). After hearing a white noise stimulus, participants were asked to input the key assigned to the discrete position at which the noise occurred, numbered 1 to 5, and presented in random order. Each sound was heard twice for each presentation condition (VBAP X LP) and with 2 repetitions for each of the 5 locations, yielding 20 trials per participant. Cofino et al. found no statistically significant result to support the hypothesis that VBAP would result in better localization compared to LP. However, there are many problems with their experimental design.

2.4.3 Wave Field Synthesis

While VBAP aims to compute a phantom image, and then apply gain manipulations to a set of loudspeakers in order to render said image, another technique is to actually synthesize the desired signal: for each point in time, compute the pressure field at a given point, which over time gives a mathematical representation of the sound wave as it propagates through a given space. Start gives a nice introduction to the concept in his PhD thesis [36]. He conceptualizes sound propagation as a series of increasingly decaying pressure points in space. At time t, a sound source emits a wave, called the fundamental or primary spherical wavefront. At time $t + \delta$, each point on the sphere is a new decayed point source of the original and thus emits a secondary wave. The envelope (or sum) of each of these secondary wavefronts is the total emanating wavefront at time $t + \delta + \gamma$. Thus, by computing the pressure at each point on the primary wavefront and taking the sum, you can compute the wavefront as it would be at time $t+\delta$. This conceptualization is called the Huygen's Principle, after the mathematician who first proposed it.

Given this principle of wave propagation, Start describes how it forms the basics of acoustic synthesis. For any point in a sound source free volume, referred to as the receiver point, the received signal can be computed if the sound pressure of the wave on the surface enclosing the source free volume is known [36]. If this source free volume is squashed to an infinite plane, then only the sound pressure at each point of the plane is needed to fully reproduce the wave from the sound source. This is the basis of the wave field synthesis (WFS) technique[2]. While conceptually easy to grasp, there are considerable practical issues with resolving the waveform as it propagates, which for years made the problem rather intractable. However, as the processing power has increased, approximations to the integral of all points on the enclosing surface of the source-free volume is achievable, which has led to interest in applying wave field synthesis to real-world applications.

Of course, the infinite plane described above is purely theoretical. In implementing WFS, we typically have a linear array of loudspeakers; a discretised method

[2]The mathematics for wave field synthesis is out of scope for this thesis. I'd like to direct interested readers to the work of Start and Berkhout et al. [36, 37].

for sound reproduction. To this end, there are mainly issues with frequency aliasing (due to the discrete points on the plane as represented by the speaker array) and truncation (due to the fact that the speaker array is of course not infinite). Having said that, there are many wave field synthesis platforms implemented today. Rébillat et. al developed a platform for audiovisual 3D display called the SMART-I^2. It renders virtual environments using a visual display similar to the CAVE (discussed in Chapter 3, Section 3.2) system, and implements a wave field synthesis sub system for audio rendering. Another system is the Soundscape Renderer (SSR) by [38]. The SSR is a flexible acoustic renderer, which implements VBAP, Binaural (See Section 2.5.3), and WFS. The SSR has been successfully used in many applications such as the BoomRoom [39], and I will discuss it in detail in Chapter 5 as it formed part of the practical element of my EngD.

The latest research in WFS focuses on interactive virtual environments. Mehra et al. developed techniques for interactive wave propagation for virtual reality applications [31]. They note the main problems with WFS is the high computational load with respect to high-frequencies, and the lack of a solution for dynamic, real-time listener directivity. Their solution lies in a hybrid pre-computational/real-time system. In the pre-computational stage, they compute the pressure field for a series of sound source based on their directivity. They break the problem into two parts: the source directivity and the listener directivity. For each source in the scene, they compute a set of surface harmonics (SHs) (See Section 2.5.6), and represent them as a set of basis functions. At runtime, given the dynamic source directivity, they decompose the SHs to a set of coefficients, and then sum the pressure field based on a weighting scheme of the coefficients. For the listener directivity, they track orientation of the listener and decompose the plan wave of the pressure field, then use this combined with the head-related-transfer function (HRTF) data to compute the responses for both ears of the listener (See Section 2.5.3). Finally, the dry audio signal is convolved with ear responses to produce the propagated wave form through the space.

Mehra et al. do provide a user evaluation of their approach to interactive wave propagation. They compare their method versus another propagation technique from their previous work [40]. They rendered videos which had an audio track rendered using the method from [40] or [31], and asked participants to rate each

video with respect to the realistic sound quality across 3 distinct scenes (3 X 2 design) based on a 2 item questionnaire: Their results showed a preference for the newer method from [31] overall, however the results of ANOVA were more interesting. There was a statistically significant main effect of render type and of the scene with respect to the first question. There was also an interaction effect between the scene type and the render type. For the second question, there was only a statistically significant effect of the render type, but no interaction for the scene. This is interesting as if the conclusion drawn is that the render method introduced in [31] is more realistic, one would expect to find an interaction effect as the visuals differed (i.e. a strong correlation between the visual and acoustic elements produced).

I interpret this as a need for more rigorous perceptual studies and empirical data with human participants, as my own work which I discuss in Chapter 5 proposes the idea of rendering *less* realistic audiovisual environments: incongruent environments, where visual and audio cues do not correlate one-to-one as they would in a physical environment. I defend this through empirical human studies data that suggests how these ***intentionally mis-mapped*** environments may be ***more perceptually veridical*** than a 1-1 mapped virtual environment. Later studies have investigated WFS further with respect to real time dynamic rendering [41], but remain out of scope for this thesis and will not be discussed further.

Having discussed some of the core techniques for rendering spatial audio, I now turn to more detail on something I touched on briefly in this section regarding WFS: namely the idea of simulating directional audio over headphones for VR. The standard technique, which I discuss in the next section, is known as binaural rendering.

2.5 Spatial Audio Production With Headphones

For decades, the dominant form of headphone listening has been stereo playback over a mobile device. However, as DSP has matured and mobile hardware has become faster, with more random access memory (RAM) capacity, it is now possible to perform complex real time processing for spatial audio rendering on a mobile

device. Binaural listening is ideal for a mobile device, as it does not require any special hardware or headphones to implement (all processing is done in software). Through standard, commercial, off-the-shelf headphones, a spatial audio experience can be delivered that avoids the need for sweet-spots and complex speaker set-ups as described in Sections 2.4.2 & 2.4.3. However, it does suffer from its own unique problems related to individual perception using generically captured models of sound perception as well as confusions between objects spatialised in front and behind the listener (discussed in Section 2.5.3).

The next section introduces the concept of binaural listening, and discusses the relevant DSP required to spatialise an audio signal over headphones. I begin with the basics of binaural (2 ear) perception of sound, discussing the time and level differences of the sound signal as it arrives at both ears, and the role this information plays in sound localization. I then shift to the processing basics of Fourier Transform and signal convolution: mathematical operations which first transform the signal into a representation that is less intensive to process in the Central Processing Unit (CPU) of the machine (Fourier Transform). Then the signal is integrated with a pre-computed 'spatialisation' signal in order to produce a third signal (convolution). This third signal is what is delivered over the headphones and gives the perception of a virtual acoustic image coming from a particular point in space.

2.5.1 Binaural Listening

Auditory localization is the ability to determine the position of a sound source with respect to the observer's current head orientation. Many studies have sought to determine the limits of auditory localization through a series of controlled experiments that modulate the position of the sound source and seek a response from the listener. Another important aspect of sound localization studies is categorising the cues to sound position; which characteristics and properties of the sound provide cues to its position, and how these cues interact with one another.

Localization can be divided into the ability to localize the position of a sound source on 2 distinct 2D planes; the azimuthal or horizontal plane, and the vertical plane. Localization is also an integration process; sound is heard binaurally, with each ear contributing individually to the process as well as interaction between the two. This is most notable in studies concerning localization blur; the resolution of auditory localization is typically between $1 \sim 2$ degrees when listening binaurally, but can increase to as much as 10 degrees for monaural localization [29, 42].

Lord Rayleigh, while discussing duplex theory in horizontal localization, describes the two main binaural cues, which we refer to today as the inter-aural level difference (ILD) and the inter-aural time difference (ITD) [43]. ILD refers to the level difference of an audio signal as perceived from one ear to the other. ITD refers to the time delay perceived between the ears. ITD generally becomes less effective as the frequency of the signal increases; typically around 1500 Hz, the ITD becomes negligible as the frequency is too high to distinguish phase differences between the waves hitting the left and right ears [12, 42]. At higher frequencies, the ILD becomes a more useful cue, as the listener's head begins to act as an acoustic block, creating an acoustic shadow on the side of the head opposite to the sound source location [42]. Figure 2-2 gives a pictorial description of the ITD and ILD elements of an auditory signal. In order to render the ITD and ILD cues accurately over headphones, there is a need to capture what is known as the head-related transfer function (HRTF), discussed in Section 2.5.3. Once captured, this HRTF is processed at run time with the dry signal to be spatialised to produce the spatial stereo signal.

2.5.2 Convolution & Fourier Transform

The process to achieve a spatialised signal is known as convolution. Convolution involves taking two time series signals and producing a third: the integral over both original functions, with one of the signals time-shifted. The operation is given in Equation 2.8.

$$f * g = \int_{-\infty}^{\infty} f(\tau)\, g(t - \tau)\, d\tau \tag{2.8}$$

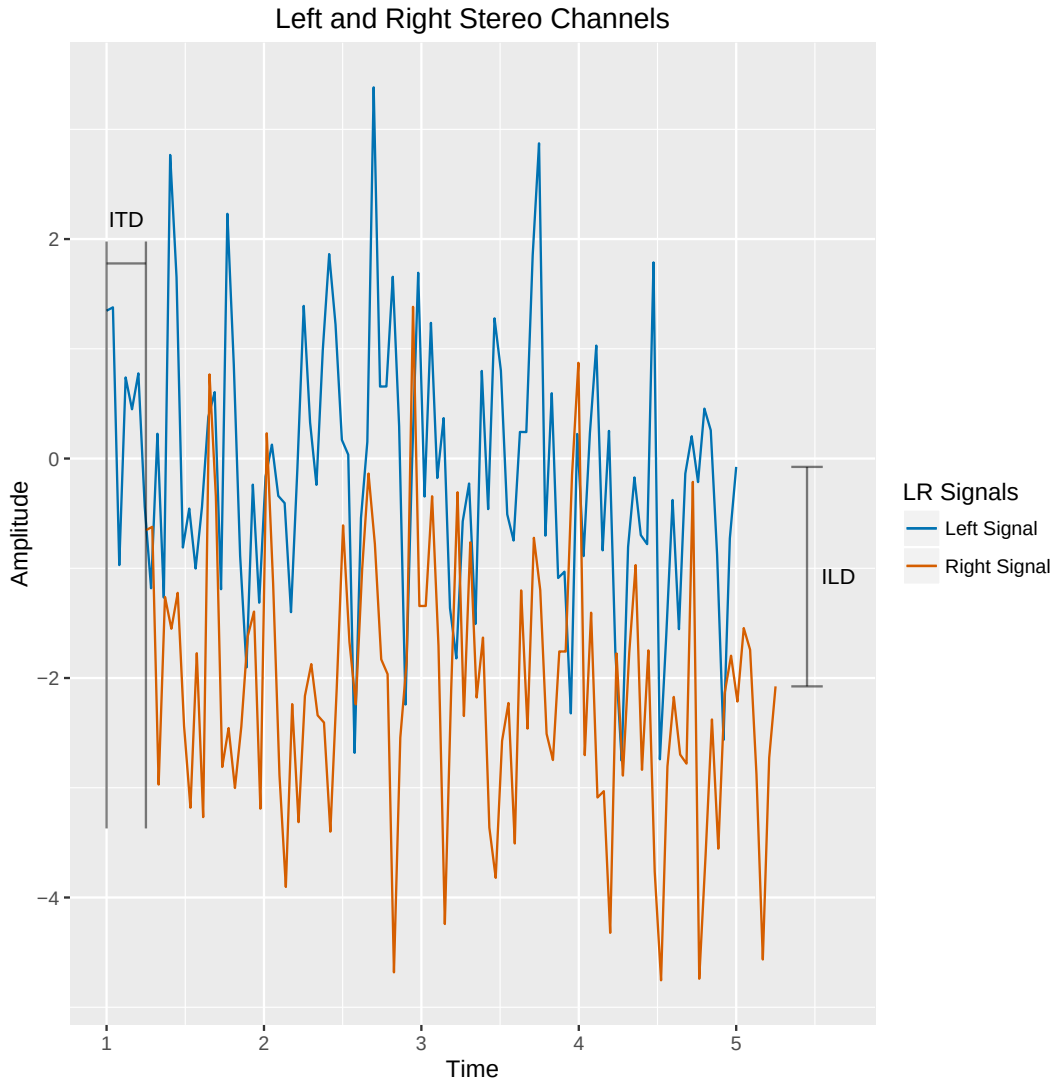

Figure 2-2: Example of the interaural time difference (ITD) and the interaural level difference (ILD) of a randomly generated auditory signal. Depending on where the sound source originates, one ear will hear the signal louder than the other, and also hear it before the other. Source code for generating image included in Appendix B.

τ represents a delay in time applied to the g signal. However, the convolution operation is much more manageable in the frequency domain, where it simply becomes a product of frequencies as shown in Equation 2.9

$$f * g = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{f\right\} \cdot \left\{g\right\}\right\} \tag{2.9}$$

This form of the convolution equation is from the convolution theorem that proves the relationship between the convolution of 2 signals and their Fourier Transforms. By taking the Fourier Transform of both signals, then multiplying them together, and then finally taking the inverse Fourier Transform of the multiplied signals, results in the time domain representation of the convolution operation. The Fourier Transform transforms a time series signal into a set of frequencies of which that signal is composed, and is typically referred to as the *frequency domain* representation of the signal. In order to convert from the time domain to the frequency domain for Fourier analysis, we apply the Fourier transform. Specifically, in digital signal processing, we apply the discrete Fourier transform. Both are shown in Equation 2.10, with the continuous form shown on top.

$$\begin{aligned} \mathcal{F}(z) &= \int_{-\infty}^{+\infty} f(x)\, e^{-2\pi i x z}\ dx,\ \ z \in \mathbb{R} \\ X_k &= \sum_{n=0}^{N-1} x_n \cdot e^{\frac{-2\pi i k n}{N}},\ \ k \in \mathbb{Z} \end{aligned} \tag{2.10}$$

The Discrete Fourier Transform (DFT) maps the signal to the frequency domain by observing that any wave form can be represented as a sum of sin and cosine waveforms. Indeed, applying Euler's formula that relates the complex exponential (the e term in Equation 2.10) to the trigonometric counterpart, makes this mapping quite clear. For completeness, Euler's formula is shown given in Equation 2.11

$$e^{ix} = \cos x + i \sin x \tag{2.11}$$

As the waveform can be approximated by this sum of trigonometric functions, the waveform can thus be represented by a sum of simpler sine waveforms [3] and it is these waveforms that the DFT extracts from the vector of time series samples. The operation essentially takes a series of samples of the audio signal, and looks for the frequencies contained in this sample. Analysing a signal in the frequency domain is convenient as the convolution theorem states the convolution between two time series signals is equivalent to the product of the DFT applied to the signals, then the Inverse DFT (IDFT) applied to the product, as shown in Equation 2.9.

[3] as a cosine wave is just a wave form with a phase shift of $\frac{\pi}{2}$.

2.5.3 Binaural Rendering: Capturing the HRTF

We can carry this analysis further to express why Fourier analysis and the convolution theorem are important to understanding the fundamentals of binaural audio. The aim of binaural processing is to reproduce the signal as it would be in the free-field or real world, coming from its given position in space, as accurately as possible over headphones. To do this, the typical method of capturing the impulse response (IR) of a set of frequencies is applied to a mannequin model of a human head. This impulse response is essentially the signal received by the microphone for a given signal sent via a loudspeaker. Wightman and Kistler detailed a process by which the audio is attenuated as it travels from the loudspeaker to the in-ear microphone [44], capturing the head-related transfer function (HRTF). Therefore, when capturing HRTF, every signal goes through the following process:

1. The signal is digitally stored in the machine as a representation of the continuous wave form.

2. This representation is then fed through the loudspeaker, which has its own IR.

3. The signal travels through the medium (typically air) by which it is altered by coming into contact with the pinnae of the mannequin model.

4. The signal is then received by the microphone inside the ear drum of the mannequin, which of course, has its own IR.

This process can be modelled as a series of *transfer functions* that modulate the signal as it passes from the loudspeaker to the microphone housed in the ear canal. Mathematically, it is a combination as shown in Equation 2.12, where the transfer functions are the Discrete Fourier Transform applied to the IRs.

$$Y_1 = X_1 * T_{Loudspeaker} * T_{Mannequin} * T_{Microphone} \tag{2.12}$$

The signal received by the microphone is a combination of the signal filtered through the response from the loudspeaker, the transfer function of the mannequin's head, and the response of the microphone itself [45]. Similarly, the mi-

crophone response is also the combination of the signal played over headphones and the transfer function of the microphone itself.

$$Y_2 = X_2 * T_{Headphones} * T_{Microphone} \tag{2.13}$$

What we want to achieve is $Y_1 = Y_2$, and thus we can solve for X_2, resulting in the transfer function of Equation 2.14.

$$X_2 = X_1 * \frac{T_{Loudspeaker} * T_{Mannequin}}{T_{Headphones}} \tag{2.14}$$

As the transfer function we are after is the change from the output signal from the loudspeaker to the output signal from a headphone, we can divide through by X_1 to get the desired complete transfer function T for synthesizing 3D sound over headphones, shown in Equation 2.15.

$$T = \frac{T_{Loudspeaker} * T_{Mannequin}}{T_{Headphones}} \tag{2.15}$$

In English terms, this means that if we have a measurement for the Loudspeaker's response, a measurement for the ear canal response, and a measurement for the HRTF, we can create a transfer function that enables the reproduction of the signal that contains the spectral information from the free-field dispersion of a sound signal from around the head. This is the basics by which binaural audio over headphones is implemented. The transfer function implements the convolution process by means of the Discrete Fourier Transform, with the response of the loudspeaker and microphone used for capturing the HRTF averaged out in the filter. Of course, this whole procedure only covers a single point source, for a single frequency, for a single ear. Thus, a complete HRTF data set generally comprises of a sinusoidal sweep at every position to capture a dynamic range of frequencies, captured twice (one for each ear) at various degrees around the head [46].

2.5.4 Evaluating Binaural Rendering Systems

Wightman & Kistler performed psychoacoustic evaluations on the binaural synthesis approach in order to test the fidelity of the virtual sound model [44]. They designed an experiment to test the localization ability of participants across free field 'natural' loudspeaker conditions compared to headphone presentation. They took absolute localization judgements in the form of a verbal spherical coordinate. In the loudspeaker conditions, the trials consisted of a series of 250ms burst of Gaussian white noise. In order to control for learned effects between the stimulus and the loudspeaker, the noise was band passed through an FIR filter mixed with random noise of a uniformly distributed random intensity. For the headphone condition, the binaural synthesis approach from [45] was applied, with individual HRTF captured for each participant in the experiment. In between trials of the loudspeaker condition, a masking stimulus was played from a loudspeaker presented above, below, in front of, or behind the participant while the experimental stimulus producing loudspeaker was moved to its new location.

In their analysis, Wightman & Kistler observed that applying traditional statistics (i.e. using the mean as a model) would be misleading. The mean azimuth error would not be justified as it is not independent of the elevation. Therefore, for analysis, Wightman & Kistler adopted a spherical statistics approach, summarising 3 separate values:

1. For every trial presented at a given $(azimuth, elevation)$ pair, they compute the angle between the unit vector from the origin to the trial coordinate and the unit vector from the origin to the response coordinate. The mean value of these angles is computed to give the *mean error angle*.

2. All judgement vectors are summed together, and then a unit length vector in the same direction is computed, called the *judgement centroid*.

3. Finally, the *index of dispersion* is computed by Equation 2.16 [4].

$$K = \frac{(N-1)^2}{N(N-R)} \tag{2.16}$$

Upon computing these values, they then took correlations between the elevation values, correlations between the azimuth values, and a goodness of fit function which applied least squares of the residuals to the target positions. Their results demonstrate a a lower mean error for stimuli presented to the sides, regardless of elevation. This may be linked to the directionality introduced through binaural listening for side stimuli. Stimuli presented on the midsagittal plane[5] are subject to the same ILDs for both ears. When presented to the side, one ear will have a greater ILD than the other. Steingrimmson & Luce discuss the notion of a bias towards the left ear in a loudness matching task [48]. In their study, participants were asked to match the loudness of a comparison tone to a standard tone, with the comparison tone played in either the right or left ear, and the standard tone played in both ears simultaneously. The important point they reference is that presentation of a sound over headphones to a single ear can lead to a localization effect towards that side of the head [48]. This is noticed in dichotic listening tests, and indeed is the basis of stereo panning [34]. Given that such a loudness bias is apparent, this may result in greater localization, as the bias itself may be providing a weighted cue to localization. This is difficult to fully appreciate in data from Wightman & Kistler given that they took the mean of the error angle across azimuth, but grouped by side, front, and back rather than left-right side [44]. The evidence for a bias is also corroborated by the review from Middlebrooks & Green, who note in the review ILDs as high as 10dB on the midsagittal plane and varying with elevation for a constant azimuth [49].

While Wightman & Kistler's study showed comparable results between free-field and headphone synthesis with respect to localization perception, they note that the process of capturing a person's HRTF is rather complex. It requires a lot of

[4] As this is an index value, I appreciate some may be interested in the full probability density function form, known as the Mises-Fisher distribution. To discuss this is outside the scope of this thesis, but I encourage the reader to See the Appendix of [44] & [47].

[5] the plane going through the centre of the human body, splitting it in two from the nose down to between the feet.

time, and requires the person to be physically present in the sound chamber in order to capture the data. This is a critical issue for applications that wish to apply a spatial audio display in an ad-hoc sense, without captured HRTF data for the individual. An important question then regarding the generalizability of HRTFs was studied by Wenzel and colleagues [25]. They found that application of generalized HRTF showed mixed results. Headphone localization by participants demonstrated comparable results with free-field listening for the majority of participants, yet some struggled with elevation localization for some of the virtual sources. However, all participants were subject to front-back reversals. The front-back reversal issue is common in HRTF based audio rendering throughout the literature. Møller reviews the aspects of binaural reproduction and notes the front-back (FB) reversal issues in binaural recordings [50]. One simple method for relieving FB reversals is dynamic movement of the head; thus this issue can be mitigated in modern systems with head tracking and real time update of the sound synthesis based on head orientation [29, 51].

2.5.5 Factors Influencing HRTF Effectiveness

More recent studies have looked deeper into the application of generalized HRTF sets on localization ability [52, 53]. Mendonça and colleagues tested the effect of training with generalized HRTF on localization ability [52]. They conducted a pilot study on the initial localization ability of participants with the KEMAR HRTF data set recording from MIT [54]. The results of the pilot suggested that the performance of participants, as measured by the percentage of correct answers, did not change over time, which led Mendonça et. al to conclude that exposure to sounds was not enough to increase localization performance. In the second experiment, a pre-test, training procedure, and a post-test phase were introduced to study the impact on localization performance. In the pre-test phase, participants were presented with each auditory stimuli, in random order, for a total of 4 times each. The training phase then incorporated a learning step where participants had 5 minutes to poll all sounds and learn their respective positions, as well as a feedback step where sounds were presented, participants had to localize them, and then they were told the correct response by the system.

This feedback step was repeated until participants could correctly localize 80% of the training sounds (trials blocked with 5 repetitions of each stimulus position). Finally, participants participated in the post-test phase. This was identical to the pre-test phase, and analysis was conducted by comparing results between the pre and post tests.

The results from the experiment were reduced localization errors (measured by angular difference between the actual rendered location and the participant response) in the post-test phase. Mendonça and colleagues then conducted another experiment, taking the same form as the one above, only modifying the elevation of the rendered stimuli and fixing the azimuth values. They found similar results to the previous experiment; training had a positive effect on the ability of participants to localize auditory stimuli presented with a generalized HRTF data set. However, It should be noted though that this experiment did not take an absolute angle measurement; participants were asked to choose the position of the stimulus from a finite set of angles.

2.5.6 Binaural Synthesis

As mentioned previously, capturing a person's HRTF can be a tedious and tricky process. While the use of non-individualized HRTF has been shown to be effective, other work has addressed the concept of partially synthesizing HRTF data. By capturing a set of HRTF at discrete locations, in order to present the listener with a sound located at a position missing from the data set, a first technique would be to interpolate between the two nearest neighbours of the new position.

Work by Evans et al. modelled HRTF as continuous functions through the use of surface spherical harmonics (SSHs) [55]. They acknowledge the traditional modelling of functions for Fourier analysis typically takes the form of trigonometric functions as basis functions (See Equations 2.10 & 2.11 in Section 2.5.2). They instead take a set of SSHs as bases, and as the SSHs are an orthogonal set (similar to a set of linear independent vectors for a given vector space $\mathbb{R}^n$), they can be used to generate any function in the space[6].

[6]The derivation of the set of SSHs is out of scope: what is important is the idea of a continuous representation for accurate interpolation between captured and non-captured IRs.

Evans et al. proceeded to capture a set of HRTF data using a custom head and torso simulator (HATS). After acquiring the data, they then analysed the HRTF using the SSH method described in the time domain, resulting in a set of coefficients for each HRTF sample. Their SSH model was then compared to the set of sampled HRTF data; for each recorded sample the error performance was taken as the root mean square (RMS) of the model recreation given the recorded sample. For `(azimuth, elevation)` pairs not captured, the model was interpolated and compared against a new set of HRTF directions not in the original set. As expected, performance was better in the recreation case than the interpolation case, yet never rose above 12%. Some shortcomings of their model is the recreation of a pre-echo effect before the actual response, due to the SSH model representation containing HRTF measured on the non-shadowed side of the HATS.

Evans et al. complete another set of analyses of their method in the frequency domain, noting again the results of a recreation comparison and an interpolated comparison. They note that the frequency domain representation resolves the issue of the pre-echo, as well as generally improving over the time domain versions for recreation comparison. With respect to interpolation, pre-echo was also eliminated, yet some artefacts still occur with regards the widening of amplitude peaks, which Evans et. al blame on the interpolation of peaks with varying numbers of samples. In conclusion, they show that their SSH model compares well with a previous model from Chen et al. [57], namely that their model is comparable across other HRTF data sets and between individual listeners, as their model is a continuous representation.

In a later study, Schissler and colleagues also applied spherical harmonics (SHs) to implement volumetric audio sources in VR [58]. They're approach can be seen as an extension to the work from Evans et al. mentioned above [55]; while originally SHs were used to interpolate through a set of HRTF for a given point source, Schissler et. al show how the same idea can be used to spatialise a set of points, representing a volume in space, and combine HRTF data from each point efficiently on modern processing hardware. Their approach is simple and

For a derivation, see [55]. For more insight into using SSH for representing HRTF data and for wave propagation, see [31, 56]

elegant; they sample the volumetric space at a series of points, then these points are projected onto a spherical model of the listener's head. This minimizes the computational load, as Schissler and colleagues note that the projected area is much smaller than the size of the actual area. They specify their model for both spherical sources and arbitrary shaped sources; the SHs for spheres can be computed analytically due to the rotational invariance of a sphere. For more complex shapes, the solution must be approximated; Schissler and colleagues apply a Monte Carlo approximation technique.

Moreover, they also ran a user study to assess the impact of their spatialization system with user's preference to the soundscape produced, directly comparing the volumetric projection method with the point-sampling method of traditional HRTF systems. Their results demonstrated a clear preference for their volumetric technique versus point sampling, across a variety of scenes in different indoor/outdoor environments. With respect to sound perception, Schissler et. al note that the SH approach produces perceptible disturbance in pure tones, due to the method of sampling. All points are assumed to travel the same distance in the projection to the listener's spherical model head, as the resulting product is a single projection point for all points on the actual object. This implies that phases shifts would be lost in the projection of large objects, as the distance between sampled points may be large enough to produce phase shift time delays (remember, all points are modelled with the same dry audio signal). However, this would only really be perceptible in pure tones, as human perception is very tolerant of phase shifts [59].

Carty details two methods for HRTF interpolation in his PhD thesis [12]. Taking a spherical model of a human head, he develops a number of distinct operations for interpolating HRTF data. The first focuses on the truncation of phase across angels of the head. At a high level, the algorithm operates by applying a 4-point linear interpolation across the HRTF data in order to compress the HRTF data into a format that is processable in real time. At a given time processing step, the current value of the phase is measured, and is truncated to its nearest neighbour. If the phase is currently in the centre between two phase values, a cross-fade occurs between the two. Carty cites Kulkarni and colleagues on the perceptual insensitivity of phase as the motivation for the phase interpolation algorithm [59].

Another method Carty details makes an assumption of the shape of the human head. Taking a spherical head model, Carty demonstrates that the ITD of the HRTF data can be computed mathematically through a simple manipulation of basic trigonometric functions. Given the time of arrival at one ear, the ITD for the opposite ear is the time taken to reach the other side. If the spherical assumption is held, then this becomes as in Equation 2.17.

$$d \sin \theta \tag{2.17}$$

Here d represents the diameter of the sphere model, and θ is the angle from the mid-midsagittal plane of the listener (i.e. their current orientation) to the sound source. However, Carty notes that the sound would not travel directly through the head of the listener, but around the arc spanned between the point where the sound hits the head, and the actual ear position. This distance depends on the angle θ and the radius of the sphere, and is shown in Equation 2.18.

$$r \sin \theta \tag{2.18}$$

He also accounts for the medium through which the sound propagates; for typical real world scenarios, this is air. As time is distance divided by speed, Equation 2.17 is extended to account for this arc and the speed in air, and becomes Equation 2.19.

$$r \frac{(\theta + \sin \theta)}{c} \tag{2.19}$$

Note that what I've discussed here would apply to azimuthal placement with respect to fixed elevation. For completion, the full elevation can be introduced by taking the cosine of the elevation angle, producing a complete model given in Equation 2.20

$$r \frac{(\theta + \sin \theta)}{c} \cos \phi \tag{2.20}$$

This spherical model is referred to as the Woodworth Model [12, 60], and has been used to develop other HRTF interpolation models using the CIPIC database from Algazi et al. [61].

Carty proposes to assess the head model with respect to empirical psychophysical data. The observation that phase is only really important for low frequency sig-

nals, and that it becomes less of a cue in high frequencies where it is ambiguous, is exploited by extracting the low frequency phase information from the HRTF data, and using it as a scaling factor to synthesize higher frequency HRTF [12]. Carty nicely finishes his novel method with a function for computing the maximum unambiguous frequency, i.e. the frequency at which above is psychophysically irrelevant, which can then be used as the threshold for empirical evaluation vs. synthesis. I have reproduced it here for convenience in Equation 2.21.

$$\frac{c}{2r\left(\theta + \sin\theta\right)\left(\cos\phi\right)} \tag{2.21}$$

Equation 2.21 represents the frequency which bisects the head in the mid-midsagittal plane; any frequency higher than this will have ambiguous phase information as it will be a multiple of 2π, which makes it difficult to distinguish between the true ITD and a phase-masked value.

2.6 Summary

In this chapter, I have discussed the fundamentals of spatial audio rendering over loudspeakers and headphones respectively. I've reviewed the various characteristics of audio signals which interact, giving us the ability to localize sound in 2D XZ planar space. I've also discussed the techniques applied to deliver a sense of space using loudspeaker arrays with various configuration for the desired fidelity of the audio field. I then focused on headphone presentation, and the binaural rendering method, how HRTF data is captured and synthesized. While this topic is vast, and has had many books compiled over the past few decades [42, 62], I have given the core information required as background to my work on distance perception in audiovisual environments.

Chapter 3

Virtual Reality, Psychophysics, and Multi-sensory Perception

"Once we can't reject technology, we'll need to explain why some fixes are better than others. If it makes us think and ask questions, it is a worthy enterprise all by itself."

Evgeny Morozov

3.1 Introduction

Virtual Reality (VR) is a catch-all term for an environment that is artificially created and immersive. There are various ways in which such an environment may be rendered. While head mounted displays (HMDs) are the most common in modern times, and the most recognisable medium for VR content delivery, it is certainly not the only one. In their early paper introducing the "Audiovisual Experience Automatic Virtual Environment)", or CAVE as it is referred to, Cruz-Neira and colleagues compare it to other virtual reality systems with respect to visual acuity and immersion [63]. Their intention was to showcase the CAVE as

more versatile than the standard cathode ray tube (CRT) and HMD technologies at the time, by discussing some of the limitations of the latter systems and how the CAVE overcame these. As I will discuss later in this thesis however, many subsequent studies in VR distance perception noted that many problems and perceptual artefacts remain in CAVEs as HMDs (read on, or jump straight to Chapter 4).

In tandem to VR hardware and software, there is another component of a VR system: namely the human user. To accommodate her, designers and system architects should be familiar with psychophysics, a term used to describe how humans perceive their environment, how they detect and analyse sensory stimuli. When simulating environments in VR, it is easy to get caught up in physically accurate simulations for rendering graphical scenes, rigid body and fluid mechanics, and acoustic propagation of sound waves. However, studying human perception of such high density sensory information may prove fruitful and help to reduce the complexity of the simulation: if humans can't perceive such physically accurate fidelity, then why bother? Also, it encourages a more holistic approach to VR design: studying cross-modal perception and the integration of multiple sensory modalities leads to an enhanced sense of presence and immersion. To ignore the auditory, olfactory, somatic, and gustatory systems would lead to a rather bland, visual-only experience.

3.2 Virtual Reality Systems

The CAVE system detailed how virtual worlds can be rendered through projecting the scene onto walls, with the observer standing in the centre of the walled environment. CAVEs may project onto all six walls (sides, ceiling, and floor), or may use a subset (e.g 4-walled CAVE) [64]. Another popular display technology for VR is the LSID, or Large Screen Immersive Display. LSIDs are large screens that incorporate hand/motion tracking, with the observer standing in front of the screen and interacting with the world using 3D goggles (for stereoscopic LSIDs) or naked eyes. LSIDs benefit from the same traits as CAVEs with respect to immersion; the entire space of the real world, including the observer's own body, is also in view. In fact, CAVEs are a type of LSID, namely a system comprising

of 3-6 screens. The typical set-up of an LSID is a single screen in front of the viewer, and may provide touch input or body tracked gesture input [65]. Single screen LSIDs are much less intrusive than HMDs, but are more restrictive than their CAVE counterparts; namely due to the fact that interaction is limited to the frontal plane, typically within the projected environment (however, some systems do permit out-of-screen interaction limited by the tracking area of the system [65]).

In contrast to CAVEs and LSIDs, HMDs are worn on the body. This instantly separates them from CAVEs and LSIDs in the sense that they are mobile. With CAVEs, observers have certain freedom of movement and orientation, while LSIDs force observers to remain oriented towards the screen. Modern HMDs are less bulky than their older counterparts, and are thus more comfortable to wear for longer periods of time. Renner et al. note that most studies with respect to distance perception have been conducted with HMDs [2]. While being mobile, HMDs do however lose the multi-user feature of LSIDs and CAVEs. This can be overcome by real time motion tracking and rendering of other user's avatars inside an individuals HMD.

VR systems can be discussed with respect to various attributes and dimensions relating to their design. Features such as field-of-view (FOV), their level of intrusion, the panoramic experience, and the representation of the body all differ from system to system. Cruz-Neira discussed the immersive domain of the CAVE with respect to these dimensions; here I discuss them across CAVEs, LSIDs, and HMDs.

FOV relates to how much of the scene a viewer can observe at any point while keeping the head stationary. CAVEs have large FOVs due to their mode of rendering–projections on to the wall–and thus do not restrict the peripheral vision of the viewer. FOV in LSIDs is the same as in CAVEs. The only impact is the size of the display, but as this is typically very large, an almost complete 170 arc can be achieved based on the distance from the screen [66, 67]. HMDs on the other hand have evolved over the past few decades, slowly increasing their FOV (compare [63], [8] [66] for FOV increase over various HMD models). Some of the problems with FOV have thus been reduced as newer systems increased their

resolution and thus FOV. Early research claimed that FOV made a significant impact on spatial perception and awareness inside HMDs, comparing the cathode ray tube (CRT) hardware at the time to poor visibility as calculated by the Snellen fraction for visual acuity to be that of a driver at night [63]. More modern HMD systems aim to calibrate the software rendering to modulate the FOV.

CAVEs are non-intrusive as they leave the viewer to move around physically, aware of what is happening in the real world and the virtual one. Contrast this with HMDs that fully restrict the viewer's FOV. By doing so, the viewer can no longer interact with the real world, and she becomes unaware of her surroundings. This is often exactly the intended effect, but causes issues when the viewer wishes to use her hands etc. Modern HMD systems tackle this problem by creating an augmented display, overlaying the view with real world information or by tracking the viewer's hands in space and rendering them inside the display [68, 69].

Panorama is an important aspect of immersion, as it creates a sense of presence due to the constant update of the display as the viewer rotates her head and body. CAVE systems provide this panorama effect easily through their on-wall projection, yet panorama is more difficult to achieve in other VR systems, and is essentially non-existent in monitor display systems, although LSIDs overcome this limitation a little by providing more context through their large surface area. Due to the single front display, LSIDs do not provide the same panoramic immersive experience as CAVEs. HMDs initially struggled with this immersive quality, due to lag in the update of the graphical display [63]. However, more modern systems with accurate motion tracking and real-time rendering have since made this less of an issue. Of course, analogue implementations of panoramas also work exceedingly well[1].

As mentioned previously, CAVEs and LSIDs are non-intrusive displays. The viewer has a physical representation of their body within in their environment– they see their *actual* body! To enable physical-virtual interaction, the system needs to track the viewer's body and map their position to the virtual world.

[1]I highly encourage the reader to visit the Panorama Racławicka in the beautiful city of Wrocław, Poland. It depicts the battle of Racławice in 1794 and is accompanied by a multilingual audio guide which gives a picture-to-timeline guide through the battle.

HMDs, by contrast, completely restrict what the viewer can see to what is rendered in the virtual environment. Therefore they must not only track the viewer but must also render the viewer's body inside the world. Studies suggest that this is application dependent however, as the inclusion of a viewer's body has shown mixed results with respect to perception inside the HMD [70, 71, 72, 73].

Typical interaction with an LSID involves hand gestures aimed in the direction of the screen itself. Like CAVEs, LSIDs do not necessarily represent the viewer's body in the system, as the viewer is not restricted in their view of their own body. This paradigm does not restrict the visibility of the viewer's arms, but does limit the movement of the viewer, as the tracking systems typically have a specified region in front of the screen within which they can sense. Many systems act along with tracking technology (vision based camera tracking or inertial sensors on the body [65]) in order to facilitate gestural input to the system. The closed-loop feedback from the real world equips the viewer with direct control over the system and visual and proprioceptive feedback of where their arms are.

As screen technology matured over the years, HCI engineers developed large screen displays. These displays would either originally consist of a system of smaller displays working together in a matrix, but soon enough the technology grew and now large, single screen displays, with refresh rates of 75Hz, now exist on the market. These screens have been used in various strands of HCI research from user interface design to computer supported co-operative work, and from visualization large scale touch input. My focus is on how these large screens, termed large screen immersive displays (LSIDs), are used in VR environments, and the actions they permit and limitations they have. In the following sections, I review LSIDs in the same vein as CAVEs, discussing them in the context of HMDs.

3.2.1 Virtual Reality as a Cross-modal Experience

From what I've introduced so far, it would be a fair assumption that all VR content is visual, and the viewer is embedded in the virtual environment by arresting their sense of sight. However, this is a very naïve assumption. To fully

immerse someone in a virtual environment, the other senses must be accounted for. Integrating touch, smell, and taste into VR has proved challenging. In their CHI 2016 workshop, Obrist et al. introduce some projects that have aimed at augmenting the tool kit of VR designers by developing interfaces for the more difficult senses [74]. The Taste+ system uses electrodes embedded into everyday eating utensils in order to augment the sense of taste through direct stimulation of the tongue while eating [75]. In addition to stimulation of the tongue, Taste+ also uses LEDs to emit adjustable light to adjust the taste sensation. This concept of taste as a multi-sensory process comes from evidence in multi-sensory perception (MSP) studies in psychology [76]. Here, variations in the colour and weight of the utensil (again, in this study a spoon was used) affected the perceived taste of the food. Taste+ applies this research from MSP in order to adjust the perceived taste by applying variable stimulation via frequency and intensity modulation of an electrical stimulation to the tongue.

Another system which aims at creating a novel interface to taste is LOLLio, a hardware and software system for controlled taste compound delivery and human computer interaction based on taste [77]. Murer et al. demonstrate their system with a hardware prototype and a software kit running on the Arduino platform [78]. While not conducting any user studies, they provide an interaction design example of LOLLio as a feedback mechanism in games. Players could be rewarded for achieving some set goal by the release of a sweet tasting product from the device. Conversely, negative behaviour in the game may be met with a nasty release of a bitter tasting product.

In the domain of smell, more prototype systems aiming to exemplify the use of smell as a HCI interface include MetaCookie+, a system based on the combination of a HMD and olfactory display [79]. MetaCookie+ operates by tracking the physical position of a cookie as it is consumed. The HMD is used to create an augmented display of the real world; MetaCookie+ comes with head mounted cameras which view the real world, then render this with the cookie overlay inside the headset. The visual overlay, combined with the olfactory display used to alter the scent of the cookie, changed its perceived taste. Narumi et al. provide results from an exploratory study using MetaCookie+. Their study involved a cookie tasting experience with a survey on whether the participant perceived a

difference in taste between an 'augmented' cookie versus a plain cookie. Note that the 'augmented' cookie was in reality identical to a normal cookie; the only difference was the use (or lack thereof) of the MetaCookie+ system. They report users perceiving a change in taste $> 80\%$ of the time.

This might seem tangential to the issue at hand, but it introduces an important topic in psychology and neuroscience that forms an integral part of this thesis, as I applied such theories in my own studies detailed in Chapters 5 & 6. The topic in question is multi-sensory integration (MSI), and details the mechanisms that describe how we perceive the world when presented with multiple streams of sensory information. Perception is of great importance to VR. As engineers are concerned with the hardware design of VR systems, HCI is concerned with the design of the interface, typically at maximising usability. Further to this, HCI is concerned with the interface design within VR environments, termed diegetic interfaces, which can effect the immersive aspects of VR [80]. However, we are not quite ready to discuss MSI in more detail until we talk about some prerequisite concepts. The following sections introduce the fundamentals that are necessary to explore MSI in more detail. I begin with a core domain in psychology that studies the brain by treating it as a kind of processing 'black box'. This domain is concerned with how we interpret signals and process information, and falls under the umbrella term of psychophysics.

3.3 Psychophysics

Psychophysics studies human response to various sensory stimuli through repetitive exposure. In psychophysical experiments, participants are subjected to trials of stimulus bursts. As a response, they are asked to change some quantified variable, discriminate between a certain pattern, or sometimes something as simple as determining whether or not a stimulus was present [81, 82, 83]. The purpose of psychophysics is to determine the change required in a stimulus in order to elicit a particular response from the participant. In this sense, psychophysics studies the perceptual system as a linear time-invariant system (LTI). An LTI is any system that maintains a linear relationship when mapping its inputs to its outputs.

In their book on extending the perceived spectrum of audio signals, a technique known as bandwidth extension, Larsen & Aarts give a introduction to signal theory [84]. For example, imagine we have a system represented as a set of inputs, an operation, and a set of outputs produced from applying the operation on the set of inputs, all taking time as a parameter and shown in Equation 3.1.

$$\hat{y}(t) = g(\hat{x}(t)) \tag{3.1}$$

where $\hat{y}$ is the output vector, $\hat{x}$ is the input vector, and g is some function applied to the input that results in the output. A system is linear if and only if each input at time t has a corresponding output at time t, the sum of its inputs equals the sum of its outputs, if the scaling of its inputs maps to the scaling of its outputs, and if it is *invariant* to time (i.e. any shift in time in the input is reflected in the output). These criteria are represented mathematically as shown in Equation 3.5.

$$y_n(t) = x_n(t), \tag{3.2}$$

$$y_n(t) = g(x_n(t) + x_{n+k}(t)), \quad k \in [0, |x|], \tag{3.3}$$

$$ay_n(t) = g(ax_n(t)), \quad a \in \mathbb{R} \tag{3.4}$$

$$\hat{y}(t - \tau) = \hat{x}(t - \tau) \tag{3.5}$$

If we assume perception operates in a linear fashion, then we would expect the these rules to apply when we pass input to an observer and ask them to respond. There is evidence to suggest that human perception and control in certain tasks are modelled by LTI systems. In their study on pilot control and human-in-the-loop closed-loop systems, Nieuwenhuizen and colleagues model the expected behaviour of pilots in a target-following task (e.g flight simulator) [85]. They designed a model system consisting of two response functions; what the human operator sees from the visual display (Input 1) and what they feel from the force feedback of the piloting controls (Input 2). The pilot is subjected to two forcing functions for the target and the disturbance of the aeroplane, and their task is to minimize the error perceived via a visual display, namely the error the target forcing function and the roll angle of the plane. The conclusion is that the LTI model results in more accurate predictions of the human operator response.

Aircraft control has also been studied with respect to multi-sensory processing for cognitive load reduction. Johnson & Dell report a study on applying 3D auditory cues in the cockpit in order to reduce the 'visual clutter' of aircraft cockpits [86]. Using a simulator, they designed a divided-attention (dual task) experiment in which participants maintained a stable position of a cross-hair over a horizon displayed on a screen. The secondary task involved participants responding to speech based requests over headphones. Eight speech based directives towards various buttons in the simulator's cockpit. The auditory directives were presented in 3 different conditions; monaural (equal volume in both headphones), stereo (maximum volume in one phone vs. silence in the other), and spatialised with the use of Head Related Transfer Functions (HRTFs; discussed in depth in Chapter 2). Johnson & Dell note that in the spatialised condition, the eight directives were presented in a spatial location that mapped to the physical location of the buttons themselves, on a horizontal plane around the participants' head. They note that all positions were kept constant with respect to the Z axis, citing issues observed in preliminary studies that suggested participants had trouble localizing in the Z plane[2], yet they provide no cited evidence of this preliminary data.

Comparing response times of the directive being raised and then acknowledged, NASA TLX scores, and the error rate of the cross-hair task, Johnson & Dell conclude that the stereo trend line of the response time variable is steeper than the spatialised one, suggesting 3D had no effect. However, it is difficult to conclude a null effect of spatialization when, as Johnson & Dell indeed point out themselves, the data were *not* statistically significant. Begault discusses the applicability of 3D audio within the cockpit at length in his book on 3D audio for VR [62]. He touches on various confounding factors persistent in the context of an audio cockpit interface. Masking of sound affecting the intelligibility of the cue, sound fidelity compared to analogue sound from the physically present co-pilot, and ambient noise from alarms all contribute to a low signal-to-noise ratio [62]. The hypothesis is that 3D audio would relieve some of this burden due to the fact that spatial location and semantic mappings have shown to be beneficial in perceptual studies [87, 88, 53]. I will discuss this in more detail in Section 3.3.3.

[2]Localization in the Z plane is the core topic of research in this thesis; more on this later.

So studying the behaviour of pilots in cockpits has been fruitful for perception research, yet psychophysical studies tend to be conducted in simpler contexts and scenarios, using various paradigms adapted to the research question. For example, investigating perceptual sensitivity to a stimulus, it may suffice to model the participants' response using a binary Yes/No paradigm. Knoblauch & Maloney give an example of such an experiment in their book on psychometrics and modelling of psychophysics data [81]. Yes/No paradigms, or Bernoulli trials, are useful in psychophysics as they can be used to plot psychometric data of a psychometric function. The psychometric function is a probability model of a humans ability to either detect a signal (declare whether or not it is present) or classify it based on its intensity (e.g louder, quieter, equal loudness). This segues us nicely into the topic of signal detection theory in the domain of psychophysics.

3.3.1 Signal Detection Theory

Knoblauch & Moloney describe the basics of signal detection theory in their book on Psychophysics in R [81]. In particular, they discuss how the typical signal detection theory (SDT) experiment is a classification task. Classification tasks are specified by their *stimulus* and *decision* dimensions. Thus, the dependent variable of any signal detection task is number of correct classifications made by the participant.

Lets run through a hypothetical example SDT experiment. In an aeroplane cockpit, the pilot is directed to perform systematic checks of remaining fuel. He is instructed to do this whenever he hears a certain tone over the headphones. The tone is masked with noise taken from a Gaussian distribution as shown in Equation 3.6.

$$f_n = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.6}$$

with variance equal to 1 and the mean at 0. We can model this task using the Yes/No paradigm from [81]. The possible responses from the pilot thus follow the standard 4-cell matrix shown in Table 3.1. The experiment is conducted by presenting a series of trials which abide by a described method. For our example, lets assume the method of constant stimuli, which dictates that trials

	Stimulus Present	Stimulus Not Present
Response Yes	True Positive	False Positive
Response No	False Negative	True Negative

Table 3.1: Potential trial responses for Yes/No paradigm.

are generated by picking from the distribution of the random variable and are temporally independent (i.e. each trial is independent of the previous one [89, 82, 90]. As noted by Soranzo & Grassi, this simplistic method for detection tasks is subject to a report bias as they are self-reported [91, 92]. Participants may become habituated in their answering; this increases the false positive/negative rates of the task. To decrease this, you may take an indirect measure such as a behavioural action; in our example, we are directing the pilot to hit perform a directive rather than hit a specific button or reply with a verbal "yes" or "no". Given the binary choice presented to the user, Yes/No experiments are also referred to as 2 Alternative-Forced Choice (2AFC) experiments [93]. Table 3.2 shows an example data sample from a single participant in the experiment.

So in each trial, the pilot is presented with an auditory stimulus drawn from a normal distribution of noise, with the stimulus either containing the signal or not. The outcome of the trial is recorded as either a 'yes' if the pilot performed

SNR	Quanta	N	Participant	Q PerCent	NumYes	NumNo
100	0	35	P1	0	0	35
113.89	0.02	35	P1	0.02	1	34
127.78	0.06	35	P1	0.06	2	33
141.67	0.08	35	P1	0.08	3	32
155.56	0.31	35	P1	0.31	11	24
169.44	0.76	35	P1	0.76	27	8
183.33	0.97	35	P1	0.97	34	1
197.22	0.97	35	P1	0.97	34	1
211.11	1	35	P1	1	35	0
225	1	35	P1	1	35	0

Table 3.2: Data from Example Experiment. For details on generation, see Appendix B.

the task and a no if they did not. The *actual* classification of the trial (signal present or not) is also recorded. At the end of the experiment, you have a matrix of responses from the participant. In the data in this example, the signal-to-noise (SNR) ratio of masking noise in the signal distribution acts as an indicator of the signal strength → high SNR results in high detectability. You then plot this data, with the SNR on the x-axis and the proportion of correct responses on the y-axis. This plot is known as the *psychometric function*. Figure 3-1 shows a hypothetical graph generated using random data in R [94, 81].

3.3.2 Interpreting Psychophysical Data

Three important characteristics of a psychophysical graph of interest are the slope and the intercept of the graph. Figure 3-1 highlights the relevant aspects; the **point of subjective equality** (PSE), the **slope**, and the **threshold** of the data. The point of subjective equality can be thought of as chance level. This is the

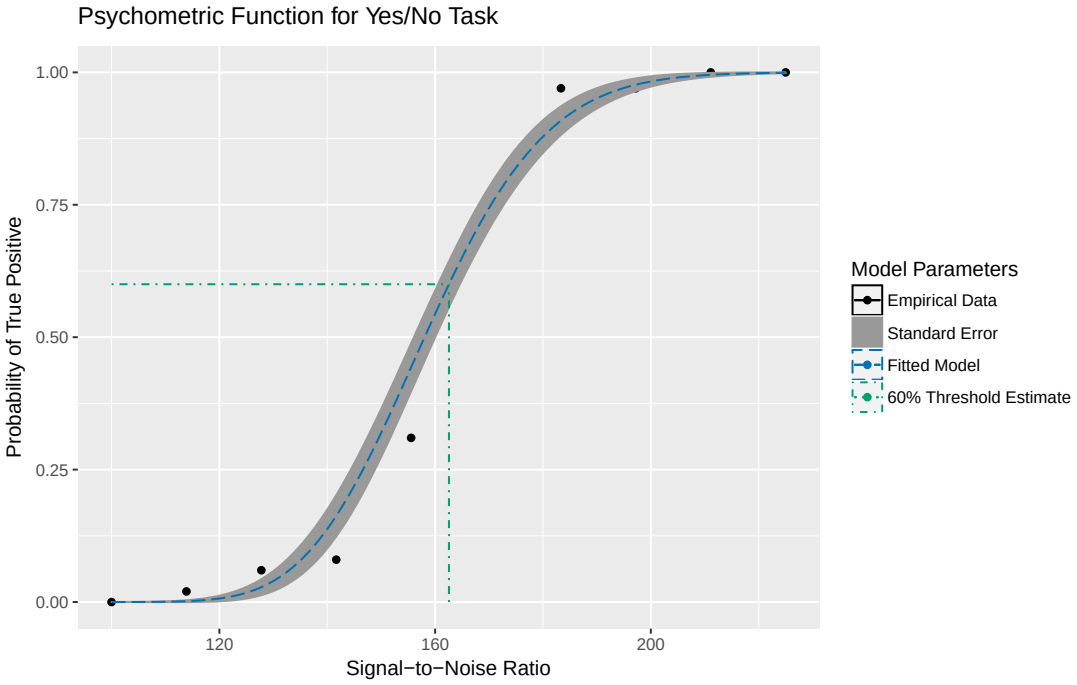

Figure 3-1: Psychometric plot for the hypothetical SDT experiment. For details on generation, see Appendix B.

noise variance level where participants basically responded 50-50 on trials, or in other words, they were more or less guessing the answer. The slope details how sensitive the participants were to the change in the variance; i.e. how quickly their performance degraded and increased with respect to the chance level. Finally the threshold is the point at which participants responded accurately enough for the task in question [95]. This is context specific to the task and is preferably determined before carrying out the experiment, or sometimes it is simply taken to be the third quantile of the data, sometimes $> 75\%$ [96].

These three values parameterise the psychometric function itself. In order to determine these values, we run a statistical analysis in the form of a regression model using a general linear model. For example, a regression model of 3 variables would be specified as shown in Equation 3.7.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{3.7}$$

This can indeed be more easily represented in matrix form as shown in Equation 3.8.

$$Y = \mathbf{X}\beta + \epsilon \tag{3.8}$$

where $\mathbf{X}$ is a matrix containing the model parameters as columns called the *design matrix* and β is a vector containing the coefficients from Equation 3.7 [81]. Psychometric functions however do not use the standard linear model, as the underlying distribution does not follow a Gaussian distribution (remember, in our example the response from the participant took the form of either 'yes' or 'no'). What to do when one does not have a normal distribution? You apply a *generalized linear model*; a linear model that may not take the Gaussian distribution predicate.

In the yes/no experiment, we have a probability outcome, plotted as a proportion of trials answered correctly. Such a distribution of scores is better represented with a Bernoulli distribution. The general linear predictor in our case is represented as a function that maps the expected outcome responses to the observed design matrix of coefficients, as shown in Equation 3.9.

$$g(E[Y]) = \mathbf{X}\beta \tag{3.9}$$

The function that makes this mapping is called the logit function [81]. In the case of a Bernoulli distribution experiment, we want to describe the function in terms of resulting in the expected value of the response being 1 (for 'yes'). This would take the form of a binomial function expressed using the logit function described above, as shown in Equation 3.10.

$$g(E[P(Y=1)]) = \log\left(\frac{p}{1-p}\right) = \mathbf{X}\beta \tag{3.10}$$

Inverting this function gives Equation 3.11, which is the mathematical representation of the *psychometric function*.

$$E[P(Y=1)] = g^{-1}(\mathbf{X}\beta) \tag{3.11}$$

In our example, the psychometric function, represented as g^{-1} in Equation 3.11 but normally referred to as ψ, tells us how the signal-to-noise ratio of the directives signal sent over headphones is mapped to a probability. Knoblauch & Maloney show how the psychometric function can be mapped to the coefficients of a general linear model by its slope and threshold values [81]. The slope and threshold values from our psychometric function are thus read as the β coefficients of a general linear model fit to the experimental data. One would like some confidence in knowing how well her data is fit by the proposed psychometric model. Wichmann & Hill give an assessment of different methods for assessing 'goodness-of-fit' [95]. They discuss the potential dangers in using standard statistical tests such as χ^2 as psychometric studies may contain small data sets. It is therefore imperative to ensure that many trial blocks are taken, or to use one of the more sophisticated methods suggested by Wichmann & Hill.

There is much more to psychophysics and signal detection theory, but that is out of scope of this thesis. The example I've demonstrated above is enough to understand the techniques applied in my studies described in Chapters 4 and 5. Now that we've covered the basics of psychophysics, I turn to areas where it is applied. The application space is vast and complex, covering aircraft control, auditory signal perception, and light sensitivity [46, 85, 97]. Psychophysical data is rich when interpreted in the context of multiple sensory modalities; signals that require not just auditory or visual processing but both together simultaneously

for example. However, this would change your model in the sense that it is difficult to distinguish the cause in the sensitivity change (slope of the psychometric function) or indeed the threshold value as there may be, albeit most definitely will be, an interaction between the two sensory modalities. Thus, we now return to discuss multi-sensory integration, a broad topic of research that aims at identifying how we process information encoded in multiple modalities and presented in multichannel form. MSI is concerned with how concurrent information streams are processed together, with their total contributions being added together in order to reach a final conclusion, a percept.

3.3.3 Multi-sensory Integration

Multi-sensory integration (MSI) is a theory of how information from various sensory input modalities is combined. The term is well defined in the neuroscience and psychology communities, where research aims to underpin the mechanisms by which we process information [87, 98, 99]. As HCI occupies itself with designing interfaces based on studies of human factors, MSI is a critical insight into designing effective models of interaction and thus merits significant attention in the HCI community. MSI is also sometimes referred to as cross-modal integration or multi-modal integration [100, 101]; both terms are used interchangeably. MSI may incorporate more than one modality at a time; it is a complex process and is not restricted to a single sense or a subset of the senses. For this thesis, I will focus on the integration of audio and visual information and show how it applies to VR. Before that, I want to give some more detail on this phenomenon now as it will make things more clear for the reader later on.

Typical MSI research presents itself in the form of a controlled experiment, where information is displayed to the participant using appropriate output displays in order to test some underlying theory of how perception works. A critical aspect of MSI is perceptual binding; the degree to which some sensory information regarding an event or object is matched with other sensory information from the *same* event or object. Of course, perceptual binding need not require multi-sensory sources of information; information streams delivered via the same sensory channel may be bound to the same event (perceptual grouping) or detached from each

other, registered as distinct percepts [102]. Spence reviewed the factors relevant to MSI of auditory and visual perception; cross-modal multi-sensory integration [87]. In his review, he details the main criteria needed for MSI to occur with respect to audition and vision. Spatial and temporal correspondence (referred to as spatio-temporal correspondence) where sensory signals are matched with respect to their source position and time of arrival, cross-modal correlation where input signals are correlated with one another on some defined dimension, are the two basic pre-requisites for MSI.

Spatio-temporal correspondence simply refers to the degree to which two sensory inputs are perceived as being sourced at the same location in space and at the same time. Studies have observed some really interesting phenomena in spatio-temporal correspondence, with the most well defined phenomena making its way into general knowledge. The Ventriloquist Effect, named for its resemblance to a ventriloquist matching the mouth of a puppet to the sound of their own voice, is however bounded by the correspondence of the two stimuli [103]. If they appear too far from one another, the effect quickly breaks down.

In an early study on the ventriloquist effect, Choe and colleagues raised the question of the definition of the effect; is it really a perceptual integration where concurrent stimuli are perceived as one, or is it merely a bias in the response from the user [104]. The point they make is that in studies looking at the after-effects of cross-modal exposure to stimuli, participants tended to bias their response in a task towards a particular sensory modality, typically vision. To give an example, consider a task that requires you to respond to a paired stimulus set where you are asked to point towards the position of the set. In a pre-trial condition, you are exposed to just the uni-sensory input, in this example an auditory beep and are asked to localize it. After this pre-test, you are presented with both the visual and the auditory stimuli together; the task remains the same. The hypothesis of the ventriloquist effect would be that your task responses would be biased towards the position of the visual stimulus. In a post-trial condition, you are then again presented with the uni-sensory input from the pre-trial condition and your responses are measured. Comparing the data in your pre and post conditions, and observing a change in your response, would be interpreted as evidence that the exposure to the visual stimulus *modified* your perception. However, the point

Choe and colleagues make is that this could be due to some learning effect or some other unknown internal bias, not necessarily a change to your perception.

To demonstrate that the former hypothesis holds (namely, that perceptual change occurs rather than a learned bias) rather than the latter, Choe and colleagues designed an experiment using signal detection theory (SDT). Their experiment consisted of a localisation task simplified to the binary model discussed earlier: participants were asked to decide whether a temporally congruent stimulus set of an audio tone and a light flash were located to the left or the right. Participants' task was to decide if the light/tone combination were presented at the same location or not. They were also asked to express a degree of confidence with regards their answer, again in binary form of sure or not sure. The proportions of responses across 4 conditions were recorded: same position and participant was sure, same position and participant was not sure, different position and the participant was sure, different position and the participant was not sure. They then formed response distributions for participants. The results from the experiment showed no difference, determined by a One-Way ANOVA on the proportions of answers across discrepancy/no discrepancy. Choe & colleagues interpret this as a sign that the underlying factor in the judgement of location is a response bias.

The Ventriloquist effect has been explained as a more general result of cross-modal 'capturing' [103]. The general concept is that certain senses play dominant roles in perception. Vision for example, with its high resolution and fast capturing speed [105], is typically the dominant sense where used, as it provides the most information. However, there is evidence to suggest that human perception *gracefully* degrades. As one sense becomes less reliable, for example walking in the dark where vision is impaired, perception becomes biased by sensory channels that are now delivering more or at least more *reliable* sources of information.

The notion of a response bias is certainly valid, yet there is overwhelming evidence for a true effect of cross-modal correspondences like the ventriloquist effect. For example, the 'McGurk Effect' demonstrates an illusion caused by the cross-modal correspondence between vision and audition, namely the impact of lip synchronisation and utterances heard [106]. McGurk and McDonald note that

the acoustic waveform between the 'ba' and 'da' utterances share similar features, while the visual lip movements of 'pa' and 'da' also share features. They give the hypothesis that these would thus be fused together, with the most common feature set, in this case that of 'da', being perceived [106]. This ***incongruity*** is interesting, as it plays a role in many other illusions. Alternatively, the effect of cross-modal congruence leading to augmented perception is observed, such as in the parchment-skin illusion described by Jousmäki & Hari [107]. By increasing the intensity of high frequencies of sound played over headphones, while simultaneously stroking the hand, participants reported an audio-tactile sensation where the skin felt drier. Even more interesting is the notion of opposite-direction effects. In the parchment-skin illusion, increasing audio frequency results in increased perceived dryness. Harrar and Spence observe a reverse direction in their studies with weight of cutlery and the density of the food [76].

While non-invasive behavioural studies are effective in observational hypothesis synthesis, they lack the clarity or insight afforded by observing events at the neuronal level. Meredith, Nemitz, and Stein aimed to specify the determinants of MSI; actually qualifying the causes of many of the cross-modal effects observed in behavioural studies [108]. Their key finding was the impact of temporal synchrony on the resulting integration across sensory information. The crucial hypothesis is not that, for MSI to occur most effectively, joint peaking of intensity is required. Rather, they report that it is the peak discharge overlap from neurons responding to the various sensory inputs that results in the highest levels of response [108]. More plus more does not necessarily equate to more. The time at which they arrive, and the temporal window spent overlapping, plays a much larger role.

3.3.4 Maximum-Likelihood Integration in the Domain of Multi-sensory Integration

Given what we've studied so far, a fair question would be 'Knowing that multiple channels are integrated unequally, is there a method for producing the *optimal* sensory integral?'. Ernst & Banks studied MSI with respect to visuotactile displays, an designed an experiment to test the hypothesis that the brain would integrate information from all relevant sensory channels to produce a percept

that is the statistically optimal possible integration [109]. They detail an experiment involving a dot stereogram and a haptic force feedback system[3], based off previous work by Heller that suggested haptic information dominates when the visual sensory signal is blurred (i.e. unreliable) [110]. Participants were then tasked to rate the tallest stimulus in a standard 2AFC task (see Section 3.3.1), similar to previous studies [95, 110, 111].

Ernst & Banks hypothesized that human performance in such tasks is optimal. Optimal performance is defined by the *Maximum-Likelihood Estimate* (MLE) given in Equation 3.12.

$$\hat{S} = \sum_{i=1}^{n} w_i \hat{S}_i, \qquad w_i = ||x_i||, \qquad x_i = \frac{1}{\sigma_i^2} \tag{3.12}$$

n is the total number of modalities involved in the percept. The equation states that each modality should be weighted by the normalized reciprocal variance (w_i is equal to taking the variance for the i^{th} modality and dividing it by the sum of the reciprocated variance of all other modalities), with the norm function fully specified in Equation 3.13.

$$||x_i|| = \frac{x_i}{\sum_{j}^{n} x_j} \tag{3.13}$$

A pictorial representative of the MLE theory is displayed in Figure 3-2. The basic idea is that, if the MLE theory of human perception were to apply, then the variance of the multi-modal percept would be reduced compared to any of the uni-modal channels involved, while the mean of the distribution would be shifted towards the uni-modal channel of least variance. In order to test this claim, Ernst & Banks plotted psychometric functions for uni-modal and multi-modal percepts of the shape discrimination task. The functions take the variance is the visual domain as a parameter (i.e variance in the visual signal). Their empirical results were in line with what MLE predicted; they concluded this as evidence that the nervous system implements an MLE integrator for multi-sensory integration. Figure 3-3 demonstrates the psychometric plot.

[3]PHANToM finger worn devices by SenseABLE Technologies©.

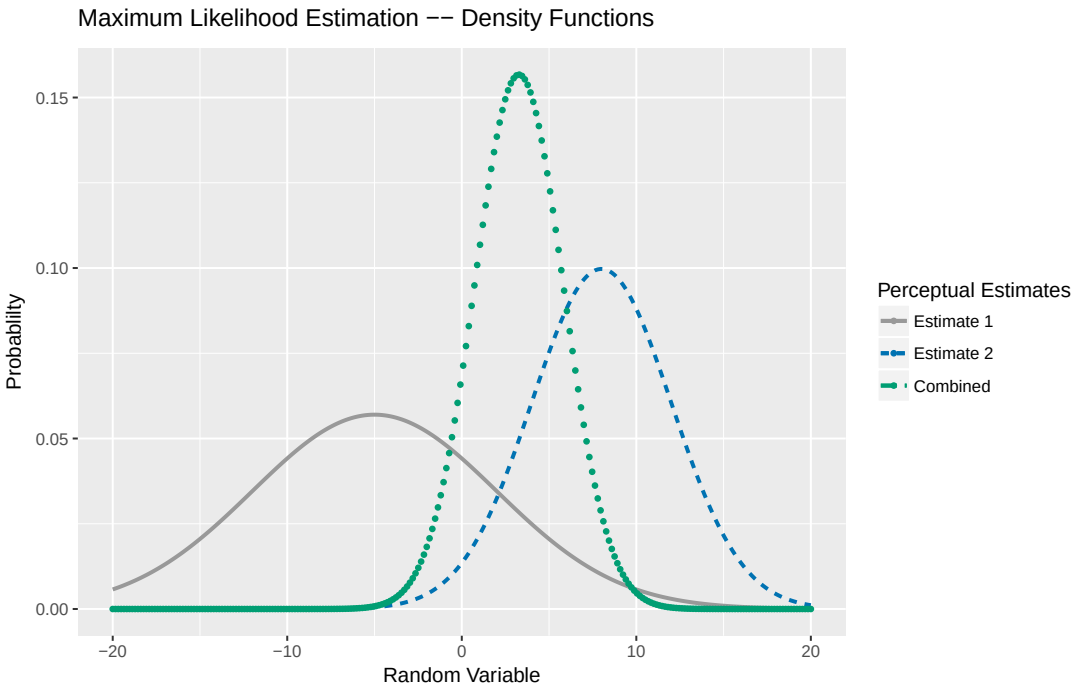

Figure 3-2: Pictorial representation of Maximum-Likelihood Estimate theory. The combined estimate from multiple modalities has lower variance than either individually, and is mean shifted towards the unimodal estimate of lowest variance. Graphs plotted using dummy data. For information regarding source code, see Appendix B.

MLE theory has been shown to also apply in audiovisual environments [103, 112, 113]. Battaglia and colleagues compared MLE to the visual capture (ventriloquist) theory and claimed both to be correct, in a sense that participants would combine cues in a way that the more reliable cue would be weighted heavier. However, they concluded that a Bayesian model which holds a prior on visual information was a better fit to their empirical data on a 2AFC localization task in the XY plane [112]. The multi-modal (audiovisual) trials had a standard stimulus that was depicted slightly offset to the left or to the right, in order to illicit the weights towards the auditory and visual modalities separately.

Alais & Burr demonstrated the reverse ventriloquist effect towards audio when visual stimuli are heavily degraded [103]. Their findings support the MLE theory from Ernst & Banks [109] by showing that the bimodal localization of an audiovisual stimulus is still better than that of either uni-modal visual or uni-modal

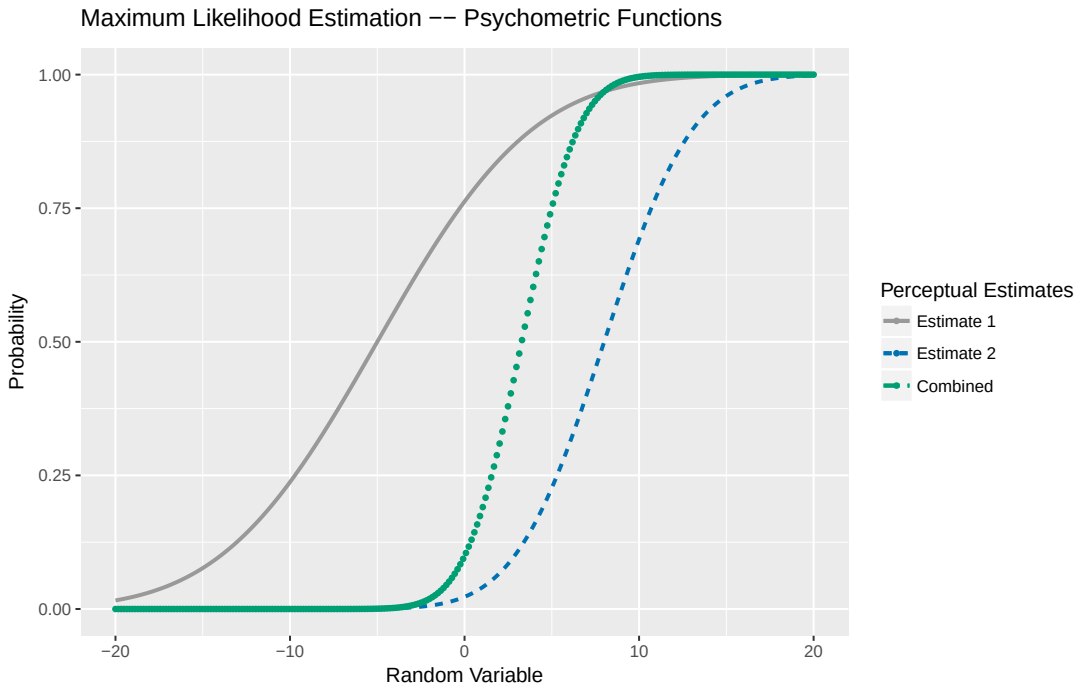

Figure 3-3: Psychometric functions (fit as normal cumulative distribution functions) of the MLE example distributions from Figure 3-2. For details on generation, see Appendix B.

auditory stimuli. They note the ecological validity of these findings as small, as an environment in the real world would never be as blurred visually for audition to truly dominate. However in Virtual Reality, this is an intriguing finding as will become clear in Chapter 4.

3.4 Summary

This chapter has introduced some of the fundamentals of perception as a prerequisite to my studies in distance perception. It began with discussing VR as a multi-sensory platform, where auditory, visual, tactile, gustatory, and olfactory channels all come together to give an enhanced sense of presence and immersion. Then, I introduced the topic of psychophysics as an approach to studying perception as a linear time invariant system, feeding inputs and analysing the impulse response. Finally, I discussed the large domain of multi-sensory integration,

where multiple sensory channels are combined to form a percept, detailing how this process is hypothesized to occur, with various empirical data from experiments in support of the hypothesis. From this, I discussed the topic of maximum likelihood integration, which simply states that the most likely answer is the one chosen when various sensory information is in conflict, or when some sensory channels involved in the integration process are no longer reliable.

This forms the basis for my work in incongruent displays and how they effectively allow for the beneficial effect of incongruence on reducing distance compression in audiovisual virtual environments. Thus far, I have discussed various topics, models, and findings in psychophysics and multi-sensory integration with respect toe audiovisual stimuli. The fact that such perceptual artefacts occur in the real world, or in experiments conducted on 2D screens in laboratories leaves a lingering question: what role does all this have to play in virtual reality? Having discussed VR from a visual perspective, as well as spatial audio, it is now possible to make the jump to the third dimension and look at multi-sensory integration in spatio-temporal localization of virtual worlds. Of particular interest to this thesis is spatio-temporal localization in the Z-plane (i.e the *distance* between the observer and objects in her environment). Indeed, the whole point of this thesis is to investigate distance perception in audiovisual virtual environments. Chapter 4 introduces the topic of distance perception in detail, and brings all what we have covered thus far into a single stream; namely the multi-sensory perception of space in VR through spatial audio and a stereoscopic visual display system, mounted in the form of a HMD.

Part II

Distance Perception & its Compression in Virtual Worlds

Chapter 4

Compression of Distance in Virtual Environments

"In these businesses, you are lucky if you don't know anything...and you fare best if you know where your ignorance lies, if you are the only one looking at the unread books, so to speak."

Nassim Nicholas Taleb

4.1 Egocentric Distance Perception

Egocentric distance perception is defined as the perception of distance between one's own self and a target, from the perspective of one's own self [2]. As virtual reality technology matures, its application domains are expected to broaden to such diverse areas as the military (e.g. remote manned drones) and emergency services training simulations. Distance perception is particularly important when VR is used to simulate real world scenarios in which an action must be done quickly and accurately, e.g. reaching for an object; jumping over an obstacle or across a ravine; moving to a target. Other applications such as virtual mu-

seum tours, actions involving reaching to nearby objects, and augmented remote medical surgery (telerobotic surgery) all require a perception of space as close as possible to the real world equivalent [114, 115, 116, 117, 118].

Distance compression–where distances are underestimated–is both a uni-modal and a cross-modal phenomenon, and has been studied extensively in both audio and visual domains. Cutting & Vishton present a taxonomy of visual distance cues in their wide survey of human distance perception [1]. Renner et al. adapted their findings and specify the main visual distance cues (See Table 4.1) with respect to a set of defined perceptual spaces [2]. They segmented the space around an observer into 3 cue regions, personal, action, and vista space. Each individual space is defined by the distance from the observer, and different cues are applicable in the various spaces. Some cues overlap, and apply in more than one cue region.

Likewise, auditory distance perception research has identified the main cues available to observers. In two extensive reviews, auditory distance cues have been categorised as static, dynamic, absolute, and relative [9, 83]. Static cues are distance cues that do not require motion of the sound source or of the listener. Dynamic cues are cues which are perceivable through an online feedback of the acoustic environment. For example, acoustic tau, the rate of change of sound level as the listener moves through space, can be an effective cue when the 'velocity of the sound is constant' [9]. Static cues on the other hand, are distance cues derived from the properties of the sound signal itself (e.g. intensity level, direct-to-reverberant energy, spectral content). Listeners focus on changes to these properties in order to estimate a change in distance in the real world. See Table 4.2 for a summary of audio cues to distance.

When generating a virtual audio source (such as the music in an MP3 player, the notification jingles on a mobile device etc.), over headphones, if a monaural source is presented to the listener, a perceptual phenomenon can occur where the sound is localised internal to the head. This is called lateralisation, and it affects not just the distance of a sound (in this case, 0 meters from the listener), but also the angular position of the sound source [62]. Duplex Theory, regarding audition with respect to both ears receiving slightly different signals, states that

in the real world, the signal that arrives at the both ears individually is different than the other [43]. These differences in the signal arriving at each ear are termed binaural differences, and are namely associated with the level difference between the two ears (inter-aural level difference or ILD), and the time or arrival or phase of the signal (inter-aural time difference ITD). Binaural cues have been studied primarily in the past to understand human sound source localization (i.e. angular position with respect to the observer), but they have been shown to act as important distance cues in human auditory distance perception studies.

Cue	***Type***	***Range***	***References***
Occlusion	Pictorial	Personal, Peripersonal, & Vista	Creem-Regehr et al. [119] Swan et al. [120]
Relative Size	Pictorial	Personal, Peripersonal, & Vista	Rolland et al. [121] Murgia & Sharkey[122]
Height in FOV	Pictorial	Peripersonal & Vista	Grechkin et al. [67] Messing & Durgin [123]
Aerial Perspective	Pictorial	Vista	Loomis et al. [51] Paquier et al. [124]
Motion Parallax	Non pictorial	Personal & Peripersonal	Norman et al. [125] Nguyen et al. [126]
Convergence & Accomodation	Non pictorial	Personal	Kelly et al. [127] Kuhl et al. [128]
Binocular Disparity	Non pictorial	Personal & Peripersonal	Patterson [129] Ponto et al. [130]

Table 4.1: Table of visual cues grouped by range and classified type, adapted from [1] and [2].

Cue	***Category***	***Range***	***References***
Intensity	Static	Personal, Peripersonal, & Vista	Wu et al. [118] Mershon & King [131]
Monaural Ampiltude Modulation	Static	Peripersonal	Kolarik et al. [9] Kim et al. [132]
Direct-to-Reverberant Energy Ratio	Static	Peripersonal	Rungta et al. [133] Zahorik [134]
Spectral Content	Static & Dynamic	Vista	Jeon et al. [32] Werner & Liebetrau [135]
Binaural Level Difference	Static & Dynamic	Personal	Kolarik et al. [9] Spagnol et al. [97]
Acoustic Tau	Dynamic	Peripersonal & Vista	Gordon et al. [136] Speigle & Loomis [137]

Table 4.2: Table of auditory cues grouped by range and classified type.

Cross-modal distance perception (i.e. perception based on stimuli across multiple senses) has been extensively researched with respect to vision and audition. Chan and colleagues investigated the effect of cross-modal congruency on participants' ability to localize stimuli in the depth plane. Participants were better at localizing stimuli when both the audible and visible paired components were presented congruently, that is collinear from the observer [138]. However, it is uncertain whether these results differ from what is known as the Ventriloquist Effect, where the source location of speech auditory stimuli is often localised to the nearest face in the visible environment [104]. Anderson & Zahorik observed that auditory distance perception was also greatly improved when congruent visual stimuli was present [139]. Given this evidence for cross-modal perception,

an interesting question to pose is how the mechanism of cross-modal perception operates. Specifically, how is vision and audition integrated to create a uniform percept? If perception is biased towards vision, how is bias mediated? As discussed in Chapter 3, Ernst & Banks demonstrated that, for combinations of visual and haptic stimuli, humans perceive in a 'statistically optimal fashion' [109]. The theory has been demonstrated with vision and audition, but Battaglia et al. note that their results do not fully rule out visual capture [112].

However, assuming that visual capture is insensitive to some bias between co-cueing stimuli, for example, visual stimuli that are misplaced or out of proportion in a scene, the perception process may be subject to slight manipulation. Such manipulation may result in improved distance perception. Some recent results attest to this [140]. The study reported in this paper aims to further investigate the validity of this claim. Before detailing the experiment, it is imperative to present more context to the reader. Thus, the following sections emphasize auditory, visual, and audiovisual distance perception in detail, before we discuss recent research into techniques which aim to reduce distance perception in virtual environments.

4.2 Perception and Compression of Distance in Virtual Environments

Distance perception in the visual domain is unarguably more widely studied than the auditory domain [6, 10, 141, 142]. With respect to distance compression in VR in particular, researchers have sought to identify the various factors involved in distance compression in order to either prevent it from occurring, or reduce the extent to which it occurs. Previous work has highlighted a perceived compression of distance in VEs [2, 143, 144]. In these studies, for a given task there is typically a discrepancy in participants' responses within the VE compared to the real world. When asked to make a distance estimate, people typically provide varying estimates, under the same conditions, in virtual and real environments. While research shows that individual differences do exist for distance compression, it remains a general phenomenon across the population [145].

Early research from Loomis et al. demonstrate how visual distance was compressed, with recurring egocentric depth intervals being lengthened in order to appear perceptually equal [6]. A more interesting finding was that, when participants were asked to close their eyes and walk towards a previously seen target (a so called visually-directed action), participants were able to walk to a previously seen target, while blindfolded, with relative ease, demonstrated by the fact that physical distance to the target fit linearly with blindfolded walked distance [6]. This accuracy was also found in a similar task where participants deviated from the path, but were asked to point continuously to where they felt the visual target was. Thus, visually-directed action tasks can be an accurate metric of distance perception, and can be more accurate than verbal judgements, due to potential cognitive bias on perceptual judgements [141, 146].

Distance compression in virtual environments is often observed to be more profound than in the real world [10, 147]. There are many different factors involved in VR that have been associated with distance compression (weight and inertia, movement and optical flow, graphics fidelity, measurement method etc.). An exhaustive list is not yet known. Tables 4.3 & 4.4 show some of the most widely researched factors and relevant papers. Of the various factors known, I have categorised them as being sensory, physiological, cognitive, and external or environmental related. Factors were categorised based on the keywords of articles and the IVs manipulated in the experiments reported. In the following sections, we detail studies that have looked at these factors in order to summarise research in visual distance perception.

Factor	Category	References	Notes
Inertia	External/Environment	[Willemson et. al 2008]*	Fake helmet used which replicated moment of inertia
Angular Declination & Field of View	External/Environment	[Williams et. al 2009]* [Messing & Durgin 2005]* [Toth et. al 2015]* [Corjuiera & Oakley 2013]*** [Creem-Regehr et. al 2005]* [Piryankova et. al 2013]**	Artificially manipulated eye height. Perception based on reflections of objects in environment
Familiarity	Cognitive	[Dat Nguyen et. al 2011]* [Richardson & Waller 2008]*	Adaptation to the environment over time; trial with feedback, then without feedback
IPD	Physiological	[Renner et. al 2013b] [Kuhl 2006]*	Based on individual distance between pupils, Measured for each participant, Compared against average or general IPD

* Head Mounted Display (HMD) technology.
** LSID/CAVE system.
*** Other technology.

Table 4.3: Key factors known to impact distance perception in Virtual environments.

Sense of Presence	Cognitive	[Ries et. al 2008]* [Philips et. al 2009]*	Compared real world to virtual model to series of models, Only one model actually genuine, Abstract, non-photorealistic
Use of Blur	Sensory/Cognitive	[Held et. al 2010]***	Render scene with aperture blur, Compared algorithm prediction human perception in psychophysical experiment
Measurement Method	Cognitive	[Sahm et. al 2005]* [Loomis & Philbeck 2008]	Compare measurement protocols, Varying compression rates, Evidence for top-down vs bottom-up influences

* Head Mounted Display (HMD) technology.
** LSID/CAVE system.
*** Other technology.

Table 4.4: Key factors known to impact distance perception in Virtual environments (cont).

4.2.1 Sensory & Perceptual Factors

Researchers have spent almost two decades researching the role of visual cues specified by Cutting & Vishton in [1]. Particularly the role these cues may play in explaining the deficit in distance perception. Willemsen et al. investigated the effect of stereoscopic viewing conditions on egocentric distance perception using a HMD [148]. They varied the presentation method of stereo images: they presented a single image to the dominant eye, occluding the submissive eye with an eye patch (monocular), then presented the same image to both eye viewports (bi-ocular), and finally presented two images either offset with respect to the participant's individual inter-pupillary distance (IPD), or to a standard IPD value of the mean across the general population (binocular). Their results showed participants greatly underestimated distances among the various conditions when making judgements in a virtual environment using a head mounted display, compared to the real world. Willemsen et al. concluded that stereoscopic cues were not the cause of distance compression in VEs as participants achieved similar accuracy in the real world in the monocular or binocular conditions, and similar compression in the virtual counterparts. These results are consistent with previous findings that binocular and monocular distance perception in VEs are similar.

An important aspect of stereoscopic display devices such as HMDs is the field-of-view provided by the device. Creem-Regehr et al. investigated the effect of restricting the field of view for participants in a series of distance estimation experiments in the real world [149]. A diminished field-of-view (FOV) is expected to impact distance compression due to the reduction in peripheral stimulation, thus reducing the amount of light reaching the retina of the observer and affecting spatial judgements as a whole [146]. Creem-Regehr et al. restricted the FOV inside a HMD with custom viewing goggles. The goggles applied a 42° X 30° (horizontally and vertically respectively) FOV. Head movement was manipulated between-participants, as it is known to affect distance judgements [150].

In another study, Piryankova et al. compared various display technologies to investigate distance compression under different technological conditions; 2 out of the 3 VE technologies used in this study showed no significant influence on the

error rate compared to the real world, for a number of varying egocentric distances to the target [66]. However, the third technology, a semi-spherical LSID[1] showed a significant difference in the per cent error (correct distance judgements vs. incorrect judgements) compared to an analogous real world setting. The results across the conditions suggest that distance compression is not simply caused by hardware technology. In discussing their results, Piryankova et al. speculated that a wider field of view (FOV) and resolution provided by the LSID may in combination reduce distance compression, rather than specific hardware. Jones et al. found that a large FOV (150° x 88°), or simply stimulating the visual periphery via bright light, reduced distance compression. Their results were comparable to real world spatial perception.

4.2.2 Cognitive Factors

Cognitive factors are the result of processing the sensory information, and therefore are not only dependent on the signal/stimuli themselves, but on other, unknown internal factors and biases. Distance perception involves the integration of many distance cues, combining them in order to optimise the accuracy of the percept. However, certain factors seem to take precedence over others, demonstrating a cognitive weighting or hierarchical chain of cues to distance perception. The most widely cited factor towards accurate distance perception in virtual environments is that of immersion or presence. Ries et al. conducted an experiment to compare two competing hypotheses to accurate distance perception: presence versus spatial memory [151]. They argue that presence is the dominant cue, where people would behave in the virtual environment the same way they would behave in the real world if they feel strongly that the two environments correlate with each other. Their experiment involved 3 different models of a physical room: one model was scaled 1-1 with the physical room, one was scaled 0.91 meters smaller than the physical room, and one was scaled 1.14 meters larger than the physical room. Conducting a between-groups analysis of virtual models of a physical room, their results show that participants in the 1-1 model made judgements similar to the physical environment, where participants in the other

[1]See Chapter 3, Section 3.2

groups significantly underestimated the egocentric distance to a wall. While this evidence does support the presence hypothesis, it is difficult to state with confidence due to the small sample size (n=9 in the scaled groups and n=5 in the 1-1 group).

Perception of distance can be determined by the context and configuration of cues available to us. One interesting aspect is how our perception of distance can be altered by adding noise or restricting our field of view. Held et al. provide a detailed description of how pixel blur can be used to alter the perceived depth of an image [152]. They provided an algorithm for implementing systematic blurring in order to simulate various focal lengths, as well as a model that predicts how well the algorithm's output will match human judgement. In order to evaluate their model, they detailed a psychophysical experiment conducted. They concluded that systematic blurring–blur that is consistent with relative distances in a given scene–can affect human perception of distance in a visual scene yielding less variation in perceived distance. Contrast this with later studies in audiovisual distance perception that have shown a systematic improvement in distance judgements when stimuli are not aligned with each other (See the studies conducted as part of this thesis, Chapter 5), or at least that incongruence can be compensated for [153], or that incongruity can have an observed effect on the cue combination across multiple modalities [154].

Familiarity is another important factor with respect to absolute distance judgements. It is a top-down influence of memory on the perception of distance. Waller & Richardson observed that providing participants with a period of interaction within the VE (in their case, a task that involved walking towards a visual target, and receiving visual and auditory feedback, in the form of a bell ring, upon reaching the target) resulted in more accurate distance estimations in a direct blind walking paradigm [155]. They repeated a second experiment which controlled compared the distance measurement method (namely direct and indirect (triangulated) blind walking), and found similar results where an interaction phase positively affected distance estimations (i.e. more veridical distance estimates), (c.f. [156]).

In a follow up study Waller & Richardson control for body based cues (i.e. focus

solely on visual distance cues) in a similar virtual environment to that of their previous work [144]. The accuracy improvement in participants who received visual only cues was negligible compared to those who received both visual and body based cues. Therefore, familiarity with an environment may develop when the interaction task is immersive, involving multiple sensory modalities, integrating over multi-sensory cues. The type of interaction was ruled out in a third experiment that showed similar results of an interaction phase when the task was either goal oriented or purely exploration Compare this with the work of Dat Nguyen et al. who also applied an interaction (or adaptation as they termed it) phase [126]. Their task consisted of walking to a set of virtually rendered poles in a tunnel, asking participants to stop when they were directly in between the poles. To test the familiarity hypothesis, 2 conditions were implemented were only the size of the tunnel was scaled (large $\rightarrow$ small and small $\rightarrow$ large) and 2 conditions were both the tunnel and the poles were scaled. They found that participants performed more accurately (made accurate blind walking distance estimates) in conditions where the size of the poles did not differ from the adaptation phase, yet performed more poorly when the size of the poles was scaled. They conclude that it was participants familiarity with the poles that contributed to better performance in the judgement task. As the adaptation phase was goal oriented, and Waller & Richardson demonstrated that goal directed or non-goal directed interaction were both equally valid, the data from Nguyen and colleagues are in agreement with that of Waller & Richardson.

4.2.3 Environmental Factors

The context of the observer, including the physical aspects of the real world, such as walking space, and the HMD or other hardware used to render the virtual world itself are categorised as environmental factors. These factors also play a role in distance perception. Studies have suggested that the weight of the HMD can have an impact on the perceived difficulty of a task, and therefore, impact distance perception. Willemsen et al. tested the role of inertia of a HMD in distance estimation tasks [157]. They compared results from an experiment which investigated the whether the weight of an HMD, specifically the 'static

torque forces resulting from mass distribution near the front of the HMD [sic]' [157], was a contributing factor to distance compression in visual based virtual environments. Essentially, if the HMD is top heavy, the centre of mass will be positioned at the front, forcing the participant to apply their own counter force in order to maintain a forward facing view while wearing the HMD. They tested their hypothesis by building a mock HMD, with the same inertial measurements of a real HMD and compared distance estimates made in the real world without wearing the HMD, and virtual environment with a real HMD. They found that distance in the virtual environment was compressed the most, in line with prior research, but that participants in the mock HMD condition also compressed distance. Thus, Willemsen and colleagues concluded that the weight of an HMD contributes to distance underestimation (however, see work from Firestone for an in depth argument against this claim [158]).

Another aspect of controversy in distance estimates is the measurement method used. Every measurement instrument carries with it some level of noise, yet with respect to a multi modal problem such as distance estimation, involving various visual, auditory, bodily, and cognitive cues, the measurement method itself should be controlled for as a factor in conducting experiments. Visual directed action has been shown to be very accurate in humans when conducted in physical spaces [6], so has naturally been applied as a measurement method in virtual environments. Messing & Durgin conducted experiments with a video relay of a real environment projected into a HMD viewport, and found that distances were underestimated in directed walking tasks [123]. However, their set-up involved a camera mounted off misaligned to the participant's actual eye location, and would have added weight to the system (c.f. [157]). Piryankova and colleagues discuss how action based versus cognitive based (i.e. walking as opposed to verbal reporting) methods can differ, with more compression apparent in the cognitive based tasks than the verbal ones [66]. Sahm and colleagues compared two action based tasks, namely directed walking and throwing an object towards the target, in distance estimates in virtual environments [147]. They found that both methods resulted in similar compression, and concluded that distance compression in VR is independent of the measurement method used. However, Klein and colleagues found significant differences between a a triangulated walking task, and directed and verbal tasks

when conducting experiments in wall and CAVE systems [159]. For details on CAVE systems, see [63].

The impact of a virtual environment's contents on distance estimation remains inconclusive. Mohler et al. observed how the inclusion of a self-avatar can lead to reduced distance compression in VEs [73]. McManus and colleagues showed how observing a virtual avatar performing a task, and then performing that same task oneself, with or without visual feedback of one's own actions (provided via a the use of a mirror in the VE) can have interesting results for distance estimation [70]. They conducted 3 tasks inside their environment, but we discuss only the relevant distance estimation task here. Participants who were in the self avatar condition could view themselves via the mirror as long as they liked before beginning a blind walking task. They found no statistically significant findings, regardless of whether or not the self avatar was present in the environment. Lin, Reiser, & Boddenheimer demonstrated significant effects of a self-avatar when making affordance judgements related to depth (specifically, their task involved user's judging the height of a ledge from the ground) [72]. They do however note some potential confounding factors in their study, such as the fear of standing on a ledge (See work from Stefanucci et al. [160]).

4.2.4 Physiological

Human physiology has been studied with respect to distance perception. The most obvious physiological factor is one pertinent to the position of the eyes. Specifically, the distance between each eye, namely the inter-pupillary distance (IPD), creates a stereo baseline which affects distance perception in the real world. In VR, researchers have manipulated the IPD by calibrating the HMD to varying IPD values. For example, Renner et al. adjusted the IPD in a reaching task and found that adjustment of the stereo base reduced distance estimate errors, but had the side effect of negatively affecting the size perception of objects in the environment [115].

The height of the observer is a closely related factor to distance perception. In particular, the floor to eye level height directly affects the angle of declination (AoD). The AoD is a theory of distance perception, that relates the perceived distance to an object with the angle between the eye line (line of sight to a target) and the eye horizon (eyes looking straight ahead, parallel to the floor). Ooi and colleagues formed the theory from observing distance judgements in the real world, using prisms to increase the AoD. They demonstrated reduced underestimation open-loop, visually directed tasks (blindfolded walking to target), as well as adaptation of the visual system to the prism viewport [161]. The visually directed task was controlled for by repeating the prism adaptation step using a different task, namely throwing bean bags to a target.

Now, if eye height can indeed affect the perceived distance to an object, one might assume that children (having an eye height significantly lower than that of an adult) would overestimate the distance to targets compared to adults. However, this would be a difficult experiment to design, as spatial perception, and invariably distance perception is known to be influenced by previous experience [162], and thus adults would have more prior knowledge than young children. One way to control for this factor is to manipulate the eye height across samples of the population virtually. Corujeira & Oakley found significant interaction effects between eye height and distance, as well as an effect of eye height in making distance judgements [163]. Their design restricted participants to binomial eye height distribution: 110cm from ground level or 20cm from ground level. This corroborates the theory of Ooi & He [161], and similar effects of manipulating eye height have been shown in most recent research using action based measurement methods [164].

One important point to note, mentioned previously, is that compression is not only a problem of visual perception. Many studies have demonstrated the phenomenon in non-visual environments, mostly involving tactile and auditory displays. While tactile is out of the scope of this thesis (Chapter 5 introduces a novel correction technique for distance compression in audiovisual environments), in the next section I discuss studies observing distance compression in auditory-only environments mainly implemented using the virtual spatial audio techniques discussed in Chapter 2.

4.3 Auditory Distance Perception

Throughout the 20th century, distance perception was studied mainly by psychologists trying to understand how the human auditory system (HAS) operates. As sound travels through a medium, it is attenuated (known as dampening), where the intensity of the sound source diminishes. The extent to which dampening attenuates the sound signal has been empirically evaluated, and has been described as the inverse-square law. The inverse-square law states that, for anechoic environments, intensity is attenuated by 6 dB for each doubling of distance between the sound source and the listener [165]. Early work from von Békésy showed how the low frequency energy of a sound source can be a cue to distance, as nearby sources seemed to 'approach' the listener when their spectral signature was modified to introduce more low frequency spectra, and was reviewed by Coleman [165]. von Békésy developed his theory of how the perceived distance to an audible source may change as the source elevates in an arc over and around the listener's head, due to reverberation and masking of the incident sound wave [166].

Auditory distance cues can be classified as static and dynamic cues. Static cues typically represent properties of the audio signal itself, either in isolation or combined in different ways. Dynamic cues relate to how the sound signal changes over time. As an emanating sound source moves through space, the signal we hear is modified by the source position, and the context of the sound (i.e. room dampening, sound medium, occluding objects). As we move our head, we can also change the binaural cues that we perceive, in turn integrating over space, creating dynamic cues. Research into dynamic auditory distance perception has shown evidence for high level processing of such cues, with top down concepts such as familiarity and context influencing our perception of space and distance as discussed for visual distance perception in Section 4.2.2. In order to clarify the role static and dynamic cues play in distance perception, and the contribution they make in integrated perception, the next few sections discuss this static/dynamic dichotomy, as well as neuroscience research done in this area.

4.3.1 Static cues

The human auditory system integrates the different properties of a sound signal in order to make a distance percept, among others such as signal discrimination, spatial location, and recognition. The most recognizable property of a sound signal is the level or intensity of the sound. Intensity is an excellent relative cue to distance. Human ability to discriminate between intensity differences has been shown to be quite accurate [167], and the results from Mershon & King are in agreement.

Later studies began to determine the exact effect of changes to the acoustic environment on the auditory signal, which in turn affects the perception of the sound distance. Bronkhorst & Houtgast derived a model for predicting the perceptual judgements of distance inside various rooms [168]. Their model demonstrates a proportional relationship between the reverberant and direct energy ratios within the acoustic environment, and takes the directivity of the sound source and the volume of the space into account. They used their model to accurately predict the results of two listening experiments, showing how subjective listening experiments can be modelled mathematically, with their model accurately fitting empirical data with distances from the listener reaching up to 8 meters. By applying this model, Bronkhorst & Houtgast concluded that humans use two explicit cues to auditory distance, the direct and reverberant energy of the signal, and more implicit cues such as room volume and source directivity.

As a sound signal is composed of a set of frequencies, sound source spectra acts as another auditory distance cue. Evidence to support this claim comes from studies which have observed the impact of high-frequency to low-frequency content ratio on distance perception, for both static and dynamic sources [46, 169, 170, 136]. Butler, Levy, & Neff trialled participants in an anechoic environment, playing back recorded sounds monaurally and binaurally over headphones. Participants were seated in a stationary position, and all recordings were passed through high-pass and low-pass frequency filters in order to specify the spectral content of the signal. The trials with signals filtered to allow low frequency signals were judged to be 'further removed than high-pass sounds recorded in the same setting' [169]. Gordon et al. observed shorter judgements in a time-to-arrival (TTA) task when

participants were presented with band-pass signals in the 2000 $\sim$ 7500 Hz range compared to lower frequency bands. Wilkie & Stockman found similar results in a study on looming stimuli. Participants underestimated approaching audiovisual and audio-only stimuli greater than for visual only stimuli [23].

The energy in reverberant environments lingers longer than anechoic due to the reflections or echoes around the environment. The human auditory system is believed to integrate the intensity level of a sound source, with the energy in it's reflections, in order to estimate the distance to the sound source [83]. Reverberation as a distance cue can aid humans in making absolute distance judgements, providing the ratio between the original sound wave and the reflections remains low [134]. Zahorik conducted an analysis of how humans weigh intensity and direct-to-Reverberant (D-R) energy ratio cues when determining distance [171]. He found that the weight attributed to each cue when making distance judgements changed as a function of sound source type as well as sound source position [171]. When sounds were presented directly in front of participants, D-R was weighted more than intensity, meaning participants relied on the reverberant energy in the environment more than intensity of the direct wave to make judgements. As sounds were moved further to the lateral side of the listener, participants began to weigh intensity less and rely more on the relationship between the intensity of the signal and the D-R ratio of the signal.

4.3.2 Dynamic Cues

Dynamic cues involve the motion of the listener, the source, and/or both. Over time, the HAS can integrate cues of the distance to the source as the signal changes due to movement. For example, when a sound source is approaching or looming towards the listener, the rate at which high frequencies increase is greater than that of low frequencies. Known as the acoustic tau, this change in frequency content is known to influence distance judgements [167, 137]. Ashmead et al. performed experiments to measure the strength of acoustic tau as a distance cue [172]. Their experiment consisted of a real sound source (loudspeaker) which produced a 1500 ms white noise sample, and a repeated measures design with two listening conditions. In the first condition, participants heard the sound while

standing still. When the stimulus disappeared, participants were instructed to walk towards the source of the sound (open loop control). In the second condition, participants walked towards the sound source while it was still making sound (closed loop control). Participants were more accurate in the second condition, with Ashmead and colleagues attributing this benefit to the feedback from the online sound source as a distance cue. They repeated their experiment with another condition in order to control for the motion itself by allowing the stationary group a second round of listening to the audio signal. The same results were obtained: participants were more accurate when moving towards the source (i.e. experienced acoustic tau as a cue).

Acoustic tau is also a viable cue in time-to-arrival (TTA) tasks. When the velocity of a moving sound source is known or can be estimated, acoustic tau and TTA can be absolute distance cues [137]. Gordon et al. experimented with manipulating the frequency content of a static sound source and found greater underestimation in high frequency bands of a sound stimuli compared to lower bands [136]. They found that high frequency bands were perceived as more urgent by listeners: they cite this as a contributing factor to a TTA task. Perceived urgency and frequency content of a signal have been shown to be highly correlated [16, 173]. Urgency is an implicit cue to distance, as studies of looming objects suggest an emotional response to objects approaching the listener, and are 'consistent with studies showing low frequency sounds being perceived as further away' [9] (in accordance with Coleman and Békésy [165, 166]).

Auditory parallax occurs when sound sources are rather close to the listener: as she moves her head, the ITD values are greater for lower frequencies than higher frequencies, due to the head distorting the initial wavefront [62]. Kim et al. have exploited auditory parallax in order to attempt to control for distance perception in virtual acoustics systems using HRTFs (See Chapter 2) [11]. Their idea is based on the grounds that parallax operates mathematically by a difference in the angle created by the source distance and the listener's ears. By synthesizing the HRTFs at both ears for the listener, the crossover point (i.e. the point at which the HRTF for one ear intersects that of the the other ear) can be controlled, and the sound source will be perceived as coming from this crossover point. Kim et al. conducted perceptual listening tests and found that perceived distance can

be increased up to about 1m. After this, compression occurs and the perceived distance and simulated distance by the model did not match.

Moreover, their results showed a statistically significant interaction effect between sound source distance and sound source direction: participants' distances were perceived greater when the sound source was rendered at 45° and 135° than when at 0° and 180°. Kim et al. interpret this as evidence that the change of inter-aural intensity and level affect the perception of distance, however another possibility could be asymmetric processing across between ears, akin to a preference to sounds arriving from the sides and the front, and a negative impact of cross perception between ears. Thus, when trying to control distance perception, the sound source direction with respect to the listener must be taken into account. Considering this with respect to the results from Zahorik regarding perceptual weighting and source direction, it remains to be shown how cues are ranked with respect to auditory scene analysis (ASA), specifically with respect to sound source distance [134].

4.3.3 Cognitive Processing and Integration

Distance perception can also result from internal processing across a number of cues in concert, rather than based solely on individual sensing. This processing extracts information from all cues by integrating them together, and exploiting characteristics of the sound signal itself. For example, speech is a well recognized auditory signal with particular expected attributes. A speech source is expected to originate from the mouth of a person, and is thus expected to travel outwards, in multiple directions (due to the propagation of sound in free space), however, it is expected to mostly propagate in a particular direction most strongly, namely towards the desired target of the speech (i.e. the listener). This expectation of the origin of the sound source also has repercussions for vision; humans would then expect to see an appropriate visual cue emanating from the same source. Therefore, the integration of visual and audible cues to distance would occur, and the final distance percept would then be made. This hypothesis is supported by the literature on the 'Ventriloquist Effect', where concurrent audiovisual stimuli presented at different locations are localized with respect to the location of the

visual stimuli [103, 174]. Familiarity has also been shown to influence perception of distance. For example, we are familiar with the intense roar of an airplane and we typically don't expect the sound it makes, however loud it may be, to be in our near vicinity. In this instance, we may even overestimate the distance between ourselves and the airplane. The influence of familiarity is evidence of top-down processing of distance compression, involving a cognitive bias in perception (See Section 4.4 as well as work from Renner, Sahm, and Kolarik et al. [2, 9, 147]).

With regards to non-perceptual factors, the underestimation of looming stimuli is understood to be rooted in primitive, evolutionary mechanisms in the brain. Some researchers theorize that underestimation developed as a response mechanism towards approaching, dangerous stimuli [160, 175]. A study by Gagnon, Stefanucci, and Siegel claimed this to be exclusive of the auditory domain (i.e. not found in vision) [176]. It is also hypothesized that the frequency content of a signal, in particular high frequency content, has a semantic association to urgency [173]. More research is needed in order to understand and identify any adaptive procedures that may occur as humans transition between various emotional states, and whether or not systematic 'profiles' exist for auditory distance perception based on emotional state.

4.3.4 Technological Factors

Similar to the design of visual based VEs, there are many technological factors to be considered in the production of virtual audio even before considering perceptual factors. In virtual acoustics, techniques such as binaural capture, where the acoustics of a room are captured with paired microphones tucked in the inner ear of a mannequin's head, enable virtual reproduction via headphones of audio signals within a particular acoustic environment. In headphones-based spatialisation, headphone response, binaural impulse capturing and processing, and the performance of the software have all been considered to impact acoustic spatial perception in virtual auditory displays [177].

Kearney et al. demonstrate evidence that higher order Ambisonics technology, a form of 3D audio that implements multiple channels by decomposing the sound

field at a specific point into spherical harmonic functions (i.e. functions defined in terms of spherical coordinates), results in similar compression to that of the real world [178]. Spatial audio has a wide variety of applications, such as an interactive display using speaker arrays to implement a spatial music mixing room [39], and has been shown to provide an immersive experience. Ambisonics decoding over speaker arrays requires a 'sweet-spot', meaning that the listener's head is required to remain fixed at an acoustically optimal position in space [179]. In order to provide for more flexible head movement (since such head movement is typically desirable in HMD-based VR applications), we chose to use binaural spatial audio operating over headphones. Through digital signal processing techniques and geometric manipulations, visual and auditory distance cues can be modified to alter the impression of the virtual space.

4.4 Cross-modal Distance Perception in Audiovisual Environments

Over time, we become familiar with the size of an object. We are able to mentally push the object back, and picture how the object diminishes with respect to distance. Thus 'familiarity' can give us a sense of the spatial relationship between objects in a scene. Familiar objects act as anchors; intermediaries for a sense of scale. Familiarity has been studied across both the visual and the auditory domains. Studies have shown how humans can adapt to unfamiliar sounds through repeated exposure to the stimulus. This adaptation has been shown to be constant across reverberant and free field environments [165, 180], suggesting an influence of familiarity or rather a 'learning factor' attributing to distance perception. Kolarik et al. report studies on distance perception using speech stimuli [9]. Evidence for familiarity based on characteristics of speech rather than the repetition of speech stimuli comes from Philbeck and Wisniewski et al. [181, 182]. In the former study, participants perception of distance was more accurate when unfamiliar stimuli were presented that contained similar characteristics (e.g. the sound of a person shouting/screaming). The latter study details an interesting EEG experiment which demonstrated various regions of the brain responding to stimuli based on familiarity of the stimuli characteristics. Wisniewski et al.

authors demonstrate that the distance of phonetically similar speech was more accurately perceived than that which was only lexically similar, hinting that the characteristics of speech–the temporal and spectral aspects that are inherent in familiar, sensible speech–give rise to familiarity as an influencing factor, and implying its importance over low level signal composition.

Much less work has been done in multi-sensory cue integration across the audio and visual modalities, with respect to distance perception. Spence discusses cross-modal perception with respect to low level and high level processing [87]. Stimuli that occur spatially and/or temporally together, as well as the strength of cross-modal grouping (i.e how strongly the modalities are linked, such as audition and vision or vision and touch) are factors in low level integration. According to Spence, at a high level more complex processes occur, where stimuli are integrated based on known semantic associations (e.g. the sound of a roaring engine and the sight of a motorbike) contribute to cross-modal binding.

Cross-modal binding has been studied to observe visual or audio bias effects on distance perception [138, 139, 178, 183]. Chan et al. took distance judgements in an audiovisual target selection task. Visual stimuli consisted of a picture of a face being lit up, while auditory stimuli consisted of a speech signal of the phrase 'Hi'. All stimuli emanated from a 2D circular array consisting of a foreground arc and a background arc. The front arc was situated 60 cm from the listener and the background arc was 120cm. They measured localisation accuracy, asking participants to choose whether an audio, visual, or audiovisual appeared in either the background or foreground arc while staring at a fixation point at the 0° position. They found an interaction effect between the audiovisual modality and the presentation angle along the arc. Repeated measures across audiovisual *incongruent* trials showed accuracy was greater in the central location than the periphery. Also, audiovisual *congruent* trials were almost as accurate in the central locations[138]. This suggests that the auditory stimulus was a distraction, or at did not result in more accuracy than the visual only condition. Participants were told to ignore the auditory stimulus in the audiovisual conditions: the results could have been different if participants were told to perceive the audiovisual stimuli as a single multi-modal stimulus, as multi-modal processing is broadly considered to be more effective than uni-modal processing (See [184] for example).

Anderson & Zahorik explored how distance perception modulates across visual, audio, and audiovisual environments [139]. They experimented with audiovisual photo-realistic stimuli in the frontal plane, and took distance judgements verbally from participants across the 3 modality factors. The expected compression of distance occurred in all 3 modalities, and in agreement with [138], found that AV and V conditions were not significantly different from one another, yet visual cues resulted in less variability across trials [139]. They conclude that the presence of visual information improves accuracy in distance perception. Note that their design consisted of a captured BRIR and an image of a loudspeaker as auditory and visual stimuli respectively. With respect to the study by Chan and colleagues, participants would have made the implicit association between the cross-modal stimuli, and bound the two together.

4.4.1 Cross-modal Binding and Incongruent Stimuli

Considering that this cross-modal binding results is stronger for spatially congruent stimuli, an interesting question extends itself: 'What are the bounds below which perceptual binding can occur with respect to distance perception?'. Gorzel et al. investigated the bounds of incongruence for a range of audiovisual stimuli [153]. Running a perceptual study involving stereoscopic imagery (viewed using polarised filter based 3D glasses and a large HDTV monitor) and Ambisonics audio rendering over headphones, they found that there were margins of misalignment for which audiovisual stimuli were perceived as consistent. Specifically, for a visual stimulus at 2m from the observer, there was a margin of 1m for 50% (chance level) of estimates were made as consistent. However, when the visual stimuli was pushed back to 8m, a margin of 3m resulted in 50% performance. These trends suggest that the margin of misalignment increases logarithmically with the egocentric distance of the visual stimulus.

In their studies of egocentric distance perception in rooms, Maempel and colleagues found that there was no interaction effect between auditory and visual stimuli with respect to either distance perception or room size [185]. Strikingly, they found no underestimation of distances (i.e. compression), instead suggesting a general underestimation of room size. Their design consisted of congruent

and incongruent presentation of audiovisual stimuli, 2 music samples (speech and music), and stereoscopic images presented using shutter based glasses and a large HD monitor. However, their measurement method is noteworthy: responses were operationalised by asking participants to make verbal reports, choosing from a range of options in a questionnaire. Participants were also allowed to familiarise themselves with the questionnaire and the virtual environment. The authors give no justification as to why participants were asked to input responses in this manner. As has been discussed previously, familiarisation with an environment has been shown to yield increased accuracy with respect to distance [144, 155], and there is much literature in the field of psychology with respect to priming people before a task (See work from Kahneman for a history of priming studies [186]). Also, they do concede that overestimation only occurred in rooms with high levels of reverberation, which is known to cause overestimation of auditory distance perception for sources placed close to the listener [168]. They discuss their results in light of the vast literature base in support of the distance compression hypothesis, citing the inadequacy of other studies in faithfully reproducing the correct acoustic cues and/or using artificial stimuli such as white noise (c.f. work from Anderson & Zahorik [139], the evidence for familiarity as an influencing factor on distance perception [9, 183] and Sections 4.2.2, 4.3, and 4.4). They do however make one interesting observation on their results: namely that the combination of large acoustic and small optical room size perceived to be smaller than the combination of a large optical and small acoustic room; more evidence for visual bias in spatial perception.

Jaekl et al. conducted a study on the effects of temporal delay in audiovisual distance perception [183]. They performed 3 experiments that demonstrated how audio signal delay paired with a preceding visual stimuli were perceived as more distant than audiovisual stimuli presented simultaneously, even though the actual distances were the same. Their designs involved involved a task irrelevant auditory stimuli in the context of a visual alignment task, with visual stimuli being presented via a pair of projectors onto a flat surface. Participants were asked to position two dot clusters side by side so that they were the same distance from the observer. Each condition included a task irrelevant sound; the sound was either paired exactly with the visual stimuli or appeared 0-100ms after the visual

sound. The results demonstrated that increased asynchrony–sounds appearing increasingly later than the visual stimuli–increased the perceived distance.

This evidence for time delay influence is linked to the superior colliculus, a region of the brain whose function is believed to be the integration of auditory and visual signals for attention and localization [183]. Of interest is the evidence cited of delays in processing observed by cell isolation and mapping of the receptive fields in animals. Meredith and colleagues made the observation that maximal cell interaction, and thus sensory information integration, occurred when the periods of highest discharge overlapped [108]. While this is a logical conclusion–maximizing throughput frequency across sensory receptors–it has interesting implications on design theory for virtual environments. As the response time of auditory and visual receptors differ, optimal sensory integration will be achieved when stimuli are presented in a fashion that such 'discharge periods' overlap in a maximal fashion. Logically, this implies that multi-sensory synchrony, i.e. delivering multi-modal stimuli, for example auditory and visual stimuli, in synch with one another, may not be the correct design to achieve optimal multi-sensory perception. This has direct implications on software systems that render virtual environments by synchronizing visual and auditory stimuli onset.

4.5 Summary

By now it should be clear to the reader that distance perception is a well researched problem within the psychology and computer science disciplines. Its most striking characteristic is that, even through two decades of focused research, the phenomenon of distance compression remains an open question: why it occurs is still not fully understood. Indeed, this thesis itself does not tackle a gargantuan question. However, as we have now explored the problem domain, by assimilating all the information from the previous three chapters, certain observations can be made:

1. Distance compression seems to be sensory *agnostic*: it occurs in both auditory and visual virtual environments.

2. Distance compression does not disappear when both senses are *combined*:

it is still observed in audiovisual virtual environments.

3. The theory of multi-sensory integration holds that humans will integrate information processed from different modalities in a *statistically optimal fashion*: we weight the information from various modalities based on their reliability (low statistical variance) and sum them together to make a multi-modal percept.

4. This integration process is not merely driven by the reliability of the cues: there is an impact of the temporal order of processing. This can be modulated through *incongruent* arrival and processing of information. Such incongruence can be temporal or spatial.

The focus of this thesis has been on studying the effect of incongruity in sensory processing, driven by incongruent audiovisual presentation of relevant distance cues. Given the evidence suggesting statistically optimal integration in perception, as well as evidence suggesting ranking order of distance cues, I set out to determine whether incongruent displays, systems which *intentionally* present distance cues in a spatially incongruent manner, would lead to reducing compression. The idea is that by presenting information that is spatially incongruent, the weights of the integration process can be directly manipulated to correct for the compression observed in distance perception. This correction is akin to introducing a counter bias in the opposite direction of the compression phenomenon; for distances that are compressed, cues are made incongruent in the hope that the reliability distribution for a sensory modality will shift, pulling the global estimate closer in line to the actual intended distance of the object in the virtual environment.

The next chapter presents a compression correction method, and discusses it in the context of other various techniques that have been employed in attempts to correct distance compression. I detail two experiments objectively evaluating the effect of incongruence which applied some of the techniques discussed in Chapter 3. The aim is to measure the impact of a spatially incongruent virtual environment on distance perception in a quantifiable manner.

Chapter 5

Compensating for Distance Compression in Audiovisual Virtual Environments

"So if you're a single guy looking for love, and you're deciding which friend to bring out on the town with you, choose the one who's pretty much like you–only slightly less desirable."

Jordan Ellenberg

Humans are susceptible to biases. While regularly regarded negatively, biases can have profound effects, sometimes beneficial. Sedikides and colleagues derived a simple model for human decision making when given multiple choices, through empirical observation. They found that, when females were asked to choose a date based on a description of a male candidate's attractive qualities, deciding between two similar candidates, A & B, resulted in a 50-50 split. However, introducing a third candidate C that shared a subset of identical qualities to candidate A, yet was lacking in another quality, resulted in the female participants choosing A disproportionately to B. Simply put, if A & C are mostly similar yet with a single distinct difference, then C's presence impacts the choice between A & B: the choice is biased towards A (See work from Sedikides et al. [187],

however note contrary evidence from Frederick et al. [188]). In this chapter, I propose a technique for compensating distance compression by introducing a bias between the perceptual systems. Before introducing the technique, I first give some background to previous attempts at reducing the level of compression in virtual environments.

5.1 Distance Cue Manipulation

Zahorik describes two important results with regard to auditory distance estimation from his experiments in source position and stimulus type [171]. First, he showed that distance compression was independent of source position and stimulus type. When presented with a noise burst and a speech signal, distance estimates were shown to follow a power function fit, compressing the distance between the observer and the stimuli. This effect was observed as the angular position of the target stimuli differed from the observer's front facing direction. In a second experiment investigating the weighting of direct-to-reverberant (D-R) ratio (i.e. the ratio between the energy in the direct signal from the source to the observer and the reflection of that signal within the environment) and intensity in making distance estimates, the weights of the two cues were 'found to change substantially as a function of source signal type, source direction, and to a lesser extent, source distance' [171]. The conclusion was that D-R ratio is most likely used by the human auditory system to indicate changes in absolute distance. Discrimination between multiple closely positioned stimuli seems to rely heavily on intensity differences [83].

Given that we have control over the distance cues we present in our VEs, we can begin to consider ways in which we may manipulate the spatial environment in order to influence the observer's perception. In binaural environments, digital signal processing provides abilities to alter the intensity, frequency and reverberation present in the audio signal. Füg et al. modified binaural distance cues to study the effect upon distance perception in a virtual reconstruction of the environment's acoustics [189]. After capturing the binaural room impulse response (BRIR), an acoustic 'signature' of the room, 2 algorithms were applied to two distinct distance cues; the initial time delay gap (ITDG) and the energy decay

curve (EDC). The ITDG is the time difference between the first direct sound and the initial reflection. The EDC is closely related to the reverberation time (RT_{60}), the time taken for the source signal to fall by 60 dB within a given environment. The algorithms applied involved direct manipulation of the ITDG and the energy remaining in the room after a set time.

Analysis of their results demonstrated no interaction effect between the stimuli, but an interaction effect across the modified and unmodified BRIRs was observed [189]. Manipulation of the binaural distance cues affected distance perception, supporting the hypothesis that distance perception may be controlled by direct manipulation of the intensity and reverberant distance cues; a controlled, *algorithmic* manipulation of distance perception in auditory environments. However, since this was a perceptual listening test consisting of auditory stimuli alone, it remains unclear how this manipulation will affect perception in a multi-sensory environment such as when using a VR HMD with audiovisual displays. See [11] for more attempts at manipulating distance perception in audio.

5.1.1 Incongruent multi-sensory environments

Through manipulation of distance cues across different modalities (in this paper, we study the audiovisual modalities), it is possible to render VEs that are *not* spatially coherent. When audio cues and visual cues are rendered intentionally misaligned to one another, we call this an *incongruent* environment. Incongruent environments can shrink and/or expand dimensions across modalities. For example, a distance of 5 meters may be represented as 5 meters visually, yet the same distance may be rendered in audio through a slight drop in intensity, intentionally ignoring physical laws regarding sound propagation in space. Conversely, an acoustic field may be mapped to a virtual visual environment that is larger or smaller than the original physical environment which it represents.

Zhou et al. incorporated 3D sound into their investigations of distance perception in incongruent augmented reality (AR) environments [190]. They focused on the intensity of a binaural source as their primary distance cue for manipulation, scaling the intensity in order to exaggerate the observer's perceived distance from

the source. Their results showed that 3D audio had a significant effect on participants' ability to distinguish the relative depth of two competing audiovisual stimuli, reporting an improvement of correct distance judgements of around 250% compared to a visual only condition. They coupled their perceptual results with a questionnaire to elicit qualitative data from the participants. The audio objectively helped the participants to discriminate more accurately yet, qualitatively, more than half the participants surveyed were unclear whether the audio aided their judgement. From a psychological viewpoint, the integration of the audio stimuli with the visual stimuli results in a better estimate, even though it seems participants were not consciously aware of the benefit of the audio stimuli.

In incongruent perception studies, Gorzel et al. presented participants with incongruent, collinear audiovisual stimuli [153]. Binocular images were taken of a range of loudspeaker positions directly in front of a reference viewing point, in order to emulate photorealism in their study. A pink noise burst was presented virtually over headphones using captured BRIRs. An experimental task asked participants to state whether the sound came from in front of, behind, or the same location as a photorealistic visual representation of a loudspeaker. Their results show that for a visual distance range of 2, 4 and 8 meters, misaligned audio was still perceived as consistent with the visual object, despite being rendered at an incongruent position. Perceptual binding (i.e. the audio and visual components of the target being perceived together as a whole object) was maintained despite the incongruence between the visual and auditory stimuli. The authors concluded that there is evidence to suggest an incongruence margin between auditory and visual stimuli exists. Within this margin, stimuli are perceived as a single target entity. Outside this margin, however, the binding of the stimuli breaks down and two distinct targets are perceived. Incongruities have been investigated by other researchers in the perception of distance. In particular, Sun et. al investigated the effect of visual and proprioceptive (in this case, the strength of effort required to move a bicycle) incongruence in a distance estimation task [191]. They demonstrated an improvement in visually specified distance estimates when the proprioceptive information was inconsistent with visual feedback provided through optic flow.

In a study of depth perception with stereoscopic TV displays, Turner et al. investigated the effect of incongruent audiovisual stimuli on distance estimation [192]. They found a significant effect of incongruent presentation of audiovisual stimuli. Participants judged a stereoscopic visual image as closer to them when a temporally coherent sound was played at a closer position over speakers which were placed physically closer to the observer. This provides evidence to suggest that incongruence between stimuli can be used to add depth to a scene, with a significant margin of incongruence where the stimuli are still integrated (or 'binded' to use the appropriate psychological terminology) as a single, multi-modal stimulus.

Contrast these results with those of Chan et. al from Chapter 4, Section 4.4. They found that a spatially incongruent auditory stimulus affected the ability of participants to localize a visual stimulus but only in the periphery, where auditory perception is known to be more accurate than vision [193]. However, it is important to note that this study was carried out in a physical environment (lights and loudspeakers as in [192]), and that the task was not to make a distance judgement. Indeed, this is noted by the authors themselves in their discussion. Thus it is interesting that in addition to the factors noted earlier, the task at hand or the context of the judgements being made may also influence distance estimation, and this may be applicable only in physical, rather than virtual, environments.

Audio can be used to add depth to a scene but more research is needed to investigate the interactions between the manipulation of individual visual and auditory distance cues in an audiovisual environment. Manipulation of these cues will lead to variance in the estimates provided by the human visual and human auditory systems (HVS and HAS) respectively. To ask participants to make a single, multi-modal distance estimate in such environments is equivalent to asking them to provide a combined estimate provided by the HVS and HAS. In order to systematically to position the components of a target object or stimulus (i.e. its audio and visual components) we need a method for computing *how far* the components should be positioned apart from each other in order to reduce perceived distance compression. By anchoring to the visual component of a stimulus, we can position the auditory component by offsetting it based on the visual component's position.

5.1.2 Incongruent Positioning

Anderson et al. investigated distance compression in virtual auditory environments. In their work they provide the following exponential function for describing the degree to which humans compress distance:

$$\hat{y} = k\phi^{\alpha} \tag{5.1}$$

where $\hat{y}$ is the perceived target position, ϕ is the actual target position, and α and k are the slope and intercept respectively [139]. If the ϕ, α, and k parameters are a good representation of distance compression in VEs, they describe mathematically an equation between the actual distance between the observer and the target, and the perceived distance between the observer and the target. Given any 3 parameters to the equation, we can solve for the fourth. If we know the perceived position of the target $\hat{y}$, the slope of the function α, and the intercept coefficient k, we can solve for the actual position of the target.

We can move the variables over the equality sign in order to compute a value for ϕ based on a given value for a perceived position $\hat{y}$. This changes the semantics of the variables a little: rather than $\hat{y}$ acting as a perceived distance or position, it now represents the *desired distance* we want the observer to perceive. α, $\hat{y}$, and k maintain their semantics from the original equation. In this study, values for α and k were taken to have the values 2.22 and 0.61 respectively, based on the work by [139].

Once this positional offset has been computed, we can pass it into the binaural system's auditory distance rendering (ADR) algorithm. Combined with the visual rendering system, we can produce an audiovisual environment that is incongruent. This is the method we propose for the systematic positioning of incongruent stimuli in order to design an audiovisual VE that takes account of humans' compression of distance. In order to derive the positioning function, we begin with the function given by [139] and expressed above in Equation 5.1. Dividing through by k and taking the inverse of the function gives us:

$$\phi = \left(\frac{\hat{y}}{k}\right)^{\frac{1}{\alpha}} \tag{5.2}$$

Using Equation 5.2, we can systematically position the audio component of a virtual object incongruently to its visual component. The next section begins discussion on two experiments that apply Equation 5.2, bringing together the concepts introduced in Chapter 3 and elaborated in Section 5.1.1. The goal is to compensate for distance compression in virtual environments the incongruent environment design.

5.2 Experiment I: Examining Incongruence

In this experiment, I assessed whether incongruence of collinear audiovisual stimuli affected distance perception in a virtual environment. The experiment was composed of conditions involving uni-sensory and multi-sensory stimuli, with the virtual environment presented using state-of-the-art HMD hardware. Previous studies have applied techniques involving absolute distance judgements, however, experiments involving verbal estimates of absolute distance judgements have shown a cognitive bias in participants' concepts of different metrics [141]. Hence I used a discrimination task, a common approach in psychophysics research. A discrimination task enables one to determine the variance attributed to the weighting of sensory stimuli on the task, essential for applying the maximum likelihood theory described earlier. The experimental procedure was designed to measure the estimates for both the auditory stimuli and the visual stimuli individually.

The experimental task was designed to capture distance estimates provided by participants within the VE. Chapter 4 detailed that the distance toward an object will be compressed. If the audio component is placed at a greater distance from the observer, but within the incongruence margin suggested by [153], the auditory sensory signal should be integrated with the visual sensory signal to produce a combined distance estimate that is closer to the intended distance.

If the respective weightings of the two individual signals can be determined, it is possible to specify the positions at which we should place the visual and audio components of an object to give the desired distance perception. Drawing on ML, one would expect the auditory modality to be weighted more as the visual signal

becomes less reliable. Thus, by artificially adding visual noise to the display, the weights applied to the auditory and visual modalities change can be observed, and thereby measure the individual sensory estimates before their integration to a combined estimate. Hence, this experiment had two distinct hypotheses, namely:

H1: Rendering the audio at an incongruent position further from the observer than the visual stimulus (IV), will result in more accurate distance perception (DV) compared to conditions where both stimuli are at the same position (congruent conditions).

H2: In incongruent conditions, an increase in visual noise within the display (IV) will lead to a shift in the sensory signal weights towards the audio modality (DV).

5.2.1 Participants, Apparatus and Design

Data were collected from 18 participants (7 of whom were female), with a mean age of 28. Participants were a mixture of postgraduate students and full time employees of a small company. None of the participants declared any hearing impairments and 4 had corrected vision (i.e. they wore glasses or contact lenses). All participants took part in this experiment on the basis of written, informed consent approved by the University of Bath's Psychology Research Ethics Committee, Reference 13-204, and they were free to opt out of the study at any time and without delay. The participants were not reimbursed with monetary payment for their time, nor did they receive course credit for their participation.

An Oculus Rift Development Kit 2 HMD was used for rendering 3D stereoscopic graphics[1], with audio rendered using a custom plugin for the Unity Game Engine[2]. A pair of Sennheiser HD201 Lightweight Over-Ear Headphones was used as the audio display device. The plugin integrates the SoundScape Renderer (SSR), a GPL licensed software implementation for binaural audio from Ahrens et al. [38], with Unity for spatial audio rendering over headphones. Each participant was seated, with their chin resting on a chin rest to prevent head movement during

[1]https://www.oculus.com/en-us/dk2/
[2]http://unity3d.com/

the trials. The machine used to simulate the VE was a MacBook Pro (13-inch, Mid 2012 model) with a 2.9GHz Intel i7 processor, 16GB RAM and an Intel HD Graphics 4000 card, running OS X Yosemite 10.10.3.

The experiment used a repeated measures design, manipulating 4 independent variables (IVs): Modality, Visual Noise, Congruence, and Target Range. The modality factor was manipulated across 3 levels: visual-only, audio-only, and audiovisual. Visual noise was implemented at 3 levels via a Gaussian blur, applied in real time to the camera view texture through a custom fragment shader written in the OpenGL Shading Language (GLSL), and applied to the camera's render callback function in Unity's rendering pipeline. Blur was implemented by a Gaussian spread over the rendered scene in each frame. This approximates a Gaussian blur by sampling the texture at each pixel and taking the average of the neighbouring pixels. This neighbouring spread was kept constant at 4 pixels to make a 9x9 grid. The blur was implemented iteratively, with the number of iterations determining the blur level. The AV1 conditions used 2 iterations, AV2 conditions 3 iterations, and AV3 conditions 5 iterations. An example of what the participants saw inside the headset is shown in Figure 5-1. The virtual environment consisted of the stimuli, a white plane acting as the floor, and a blue ceiling.

The Congruence IV determined whether auditory and visual elements of a target object were presented at the same position (congruently) or not (incongruently). In 4 conditions, the auditory stimulus was positioned the same distance from the observer as the visual stimulus. In another four conditions, the auditory stimulus was offset from the visual component by applying the positioning function derived above (see Equation 5.2). The experiment had nine conditions in total: six audiovisual (three visual noise levels x two congruent/incongruent conditions), one visual-only, and two audio-only conditions. A noise free audiovisual condition was not included as it does not allow for computing the relative weights of the auditory and visual signals using ML in order to determine their respective distance estimates. Conditions were presented in randomized order across participants in order to minimize order and training effects.

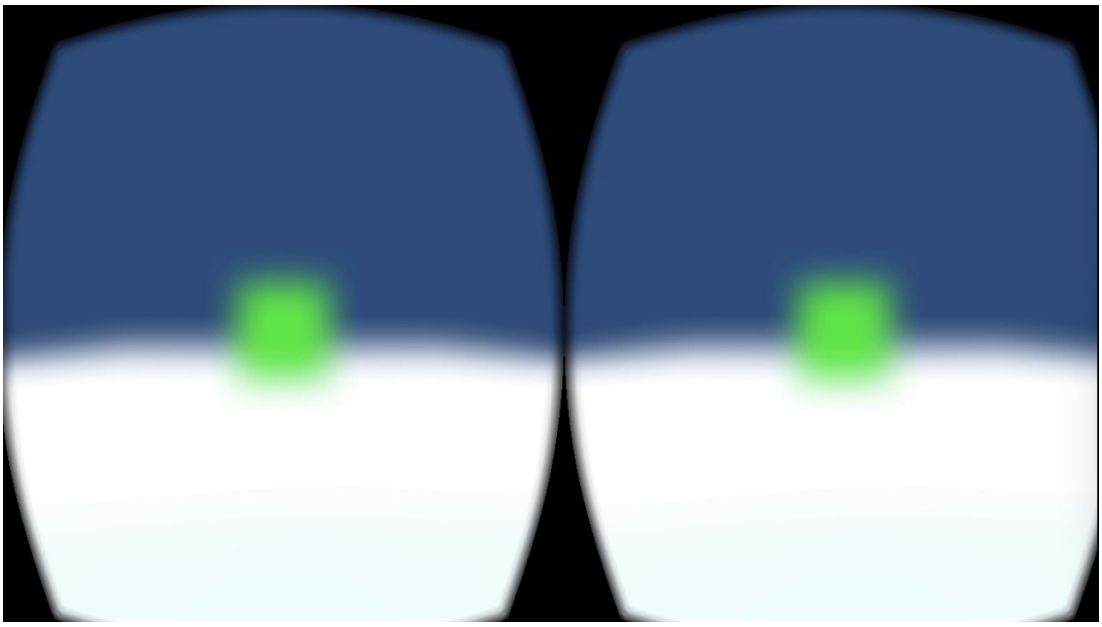

Figure 5-1: An example screenshot of an audiovisual condition in our experiment, with the visual noise at the highest level (AV3).

The range in which stimuli would appear extended to 10 meters in front of the observer, as in related work [139]. Each trial presented the target to the participant twice, with a brief (500 ms) disappearance between presentations. One of these presentations was fixed at a reference distance from the observer. The other presentation was positioned based on a staircase algorithm (see Section 5.2.2) which computed a distance offset from the reference distance. The initialising parameter of 2.5 meters, stepping down to 0.5 meters, for the staircase algorithm gave a total distance range of 0.5 to 9.5 meters in front of the observer, and split this total range into near (0.5 to 5.5 meters) and far sub-ranges (4.5 to 9.5 meters). This partition into near and far sub-ranges gave us an IV which we called the Target Range, with 2 levels. The midpoints of the near and far sub-ranges were at 3 meters and 7 meters respectively and provided the reference distances. In the congruent conditions, both the auditory and visual stimuli were presented at these reference distances. In the incongruent conditions, the visual stimulus was at the reference distance with the audio stimulus offset by the incongruent positioning function.

The stimuli presented to each participant were the same, and consisted of a visual cube, an auditory pink noise burst, or both concurrently. A pink noise burst was chosen as it distributes the same power across each octave. This avoids conflating pitch in higher octaves with magnitude [52], as frequency spectrum is known to be a distance cue [9, 83]. The distance cues available to the participant were relative size (for the visual stimulus) and intensity (for the audible stimulus).

5.2.2 Procedure

Upon entering the laboratory, participants were invited to sit down opposite the experimenter, where they were handed the HMD and asked to position it until they could see a cube clearly through the HMD viewport. The experimenter then carefully adjusted the position and tightness of the strap until the participant was comfortable, ensuring the participant's pinnae were not occluded. Next, the experimenter carefully placed the headphones over the participant's ears, and helped the participant to engage the chin rest before beginning the experimental conditions. Before commencing, all participants were subjected to an inter-pupillary distance (IPD) measurement phase. This phase calibrated the HMD for the viewer's individual IPD, and was measured using a utility packaged with the Oculus Rift SDK.

The experimental task involved, for each trial, presentation of the target (audio, visual, or audiovisual depending on the condition) at a particular distance for 500 ms. The target then disappeared and reappeared at a different distance 300 ms later. The participants' task was to indicate, using a button press on a standard computer gamepad, whether they perceived the first appearance or second appearance as closer to them. In order to choose the next distance for the target, trials followed a 3-up-1-down staircase method (See Figure 5-2). 3 correct answers resulted in reducing the relative distance between each target presentation and a single incorrect answer increased the relative distance. Guidelines from García-Pérez were followed as closely as possible in designing the staircase algorithm implemented here [194].

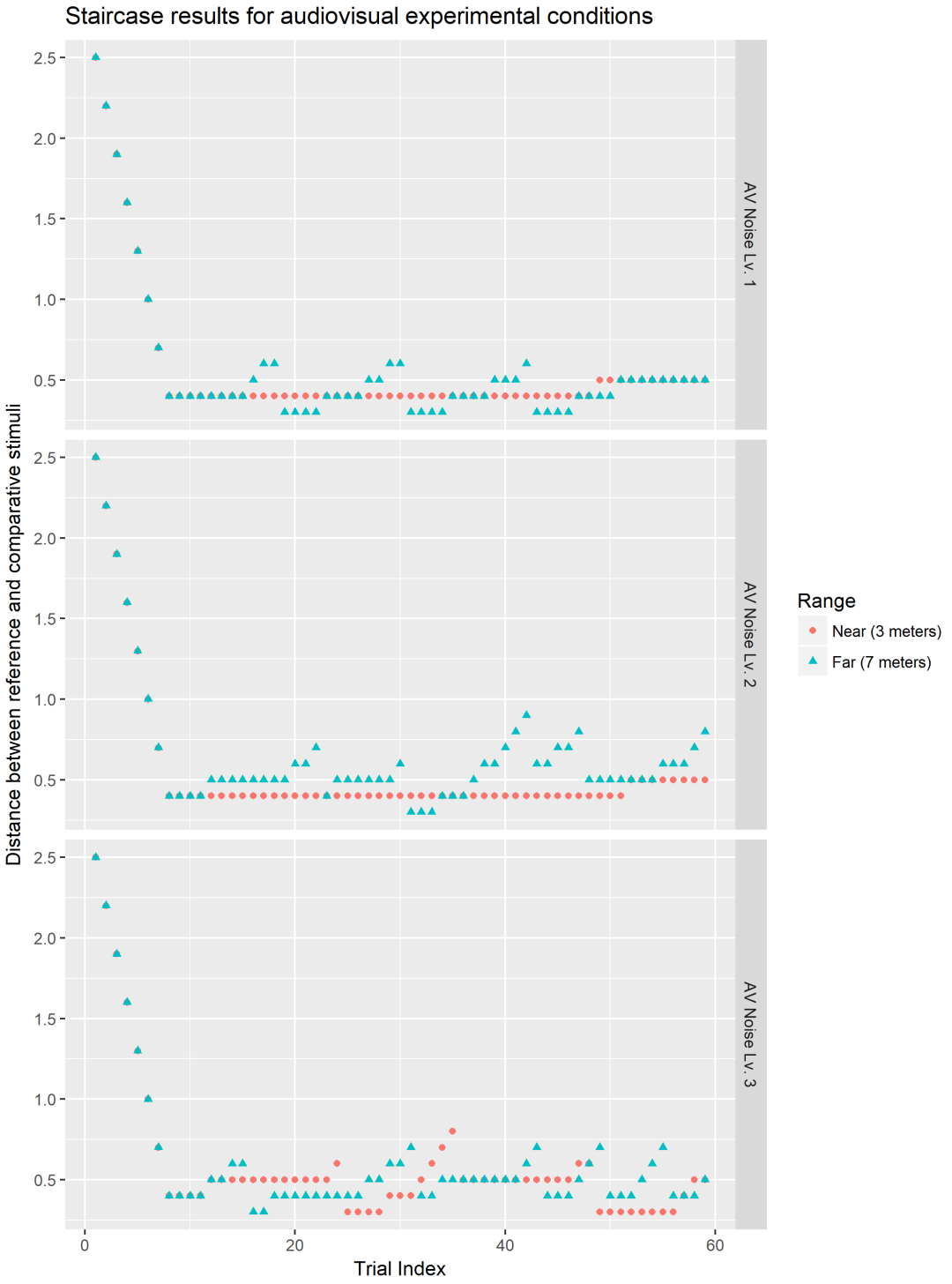

Figure 5-2: Staircase results for a single participant in an experimental session. Data are shown for all congruent conditions, in both the near and far ranges. Trials on the x-axis, distance between the stimuli on is the left y-axis, and level of noise added on the right y-axis.

Two staircases were implemented based on the Target Range, one for each sub-range. Figure 5-2 shows representative staircases, giving the results for a single participant across the AV1, AV2 and AV3 congruent conditions. Each condition consisted of 59 trials in each of the near and far sub-ranges. At the end of each condition, participants had a rest period (signalled by a red cube appearing in the centre of the display until dismissed with a double tap of the gamepad's shoulder buttons) in which they were free to remove the headset and take a break before continuing to the next condition. When they were ready to continue, they were instructed to position their head so that the environment appeared with the white plane acting as a horizon in the vertical centre of the viewport, and the red cube stimulus was directly in front of them (0° azimuth). Participants were asked to keep their eyes open during the audio-only conditions even though there were no visual stimuli in these conditions. The entire experiment took 60 ($\pm$ 15) minutes to complete.

5.2.3 Results

Data from 18 participants were evaluated, resulting in 19,116 data points across all 9 conditions of the experiment, with 118 trials for each condition, and each participant having 1062 trials. As many trials as possible were taken in order to avoid the issues noted by Wichmann & Hill (See Chapter 3, Section 3.3.2). Conditions where the audio and visual stimuli were **congruent** are termed *CON*. Conditions where the audio and visual stimuli were **incongruent**, i.e. offset with the positioning function of Equation 5.2, are termed *INCON*. All data were processed and all plots were produced using statistical packages (notably ggplot) from the R Language and Environment for Statistical Processing [94, 195].

Figure 5-3 displays psychometric function data, taking the average from 18 participants. These functions are plotted in terms of *CON* & *INCON* conditions across the three levels of the visual noise factor. All participants' results were aggregated on visual noise level, range, and distance of target. The Y-axis represents the proportion of trials where the reference interval was perceived as *closer* than the comparative interval. Also, functions are plotted with respect to the position of the visual stimulus, which acted as the anchor for the audio stimulus.

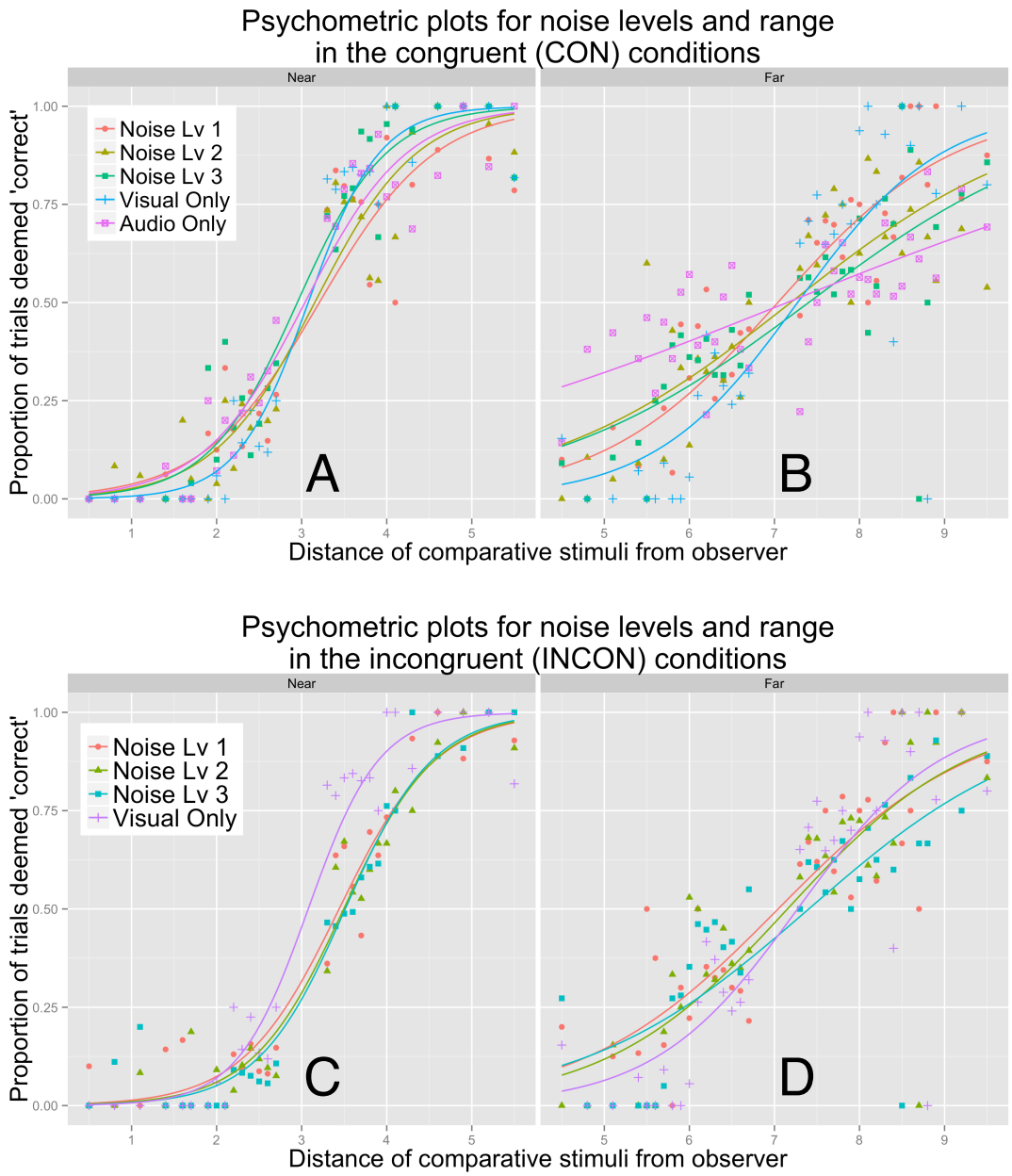

Figure 5-3: Psychometric functions in both near and far ranges, averaged over all 18 participants. Panel A shows results for the near congruent trials, panel B shows results for the far congruent trials. Panels C & D show results for the near and far incongruent trials respectively. The audio-only condition is excluded from the incongruent condition as no visual anchor was present and thus the audio stimulus cannot be 'incongruent' to a visual stimulus.

In order to plot the data, an individual trial was considered 'correct' if the participant identified the reference trial interval as being closer to the participant than the standard trial interval. ML integration weights were taken from the thresholds (at 82% correctness) of general linear model (binomial) fits to the data. The functions are as predicted from **H2**; note that the slope of the functions increase as the noise in the visual modality is increased in the *INCON* conditions. This implies that the weights have shifted to the audio stimulus, and the incongruence between the audio position and the visual position is affecting the participants' distance estimates (**H1**).

Table 5.1 shows χ^2 results for various binomial models constructed based on the distance of the target stimuli from the observer, and the effect of range on the psychometric functions. The χ^2 values indicate that the psychometric functions presented are good fits to the data. The threshold values in the table represent the distance from the observer when trials were answered correctly 82% of the time. **H1** predicts these thresholds to be reduced in the *INCON* conditions compared to the *CON* conditions.

Mean slope values for individual psychometric functions of all 18 participants were tested for the effect of incongruence. A significant effect of incongruence on the slopes of the far ranges, for all 18 participants across 6 (*CON* & *INCON*) audiovisual conditions, was observed, $t(102) = -1.84, p < 0.05, r = 0.18$. A non-significant result was obtained for the near range, $t(94) = 0.50, p = 0.69$. Thus **H1** is supported by results of the far range but not of the near range. Incongruence resulted in more accurate distance estimates when audiovisual targets were presented in the far range.

Audio modality weights in Table 5.1 were computed for the *INCON* conditions using the following equation from [109] (adapted for our experimental modalities):

$$w_A = (PSE - S_V)/(S_A - S_V)$$

where w_A is the weight with respect to the auditory modality, PSE is the point of subjective equality, or the point at which people are uncertain (chance level), and S_V and S_A are the visual and auditory estimates respectively. All conditions in the far range show a shift in weight to the audio modality ($> 80\%$ for audio),

Noise Level	Threshold	Slope	χ^2	Audio Weight
Near Congruent				
Lv 1	4.32	0.84	$\chi^2(27) = 14.037$, $p < 0.01$	N/A
Lv 2	4.13	0.84	$\chi^2(27) = 16.053$, $p < 0.01$	N/A
Lv 3	3.83	1.06	$\chi^2(27) = 18.371$, $p < 0.01$	N/A
Far Congruent				
Lv 1	8.62	0.58	$\chi^2(34) = 11.021$, $p < 0.01$	N/A
Lv 2	9.41	0.41	$\chi^2(34) = 6.311$, $p < 0.05$	N/A
Lv 3	9.71	0.40	$\chi^2(34) = 5.735$, $p < 0.05$	N/A
Near Incongruent				
Lv 1	4.39	0.95	$\chi^2(27) = 16.011$, $p < 0.01$	0.41
Lv 2	4.38	1.02	$\chi^2(27) = 16.820$, $p < 0.01$	0.39
Lv 3	4.38	1.02	$\chi^2(27) = 17.094$, $p < 0.01$	0.39
Far Incongruent				
Lv 1	8.78	0.53	$\chi^2(34) = 9.459$, $p < 0.01$	0.81
Lv 2	8.77	0.57	$\chi^2(34) = 10.580$, $p < 0.01$	0.83
Lv 3	9.40	0.46	$\chi^2(34) = 7.378$, $p < 0.01$	0.91

Table 5.1: Table of χ^2 results for the *CON* and *INCON* conditions (goodness of fit) shown in Figure 5-3. Weights were computed for the *INCON* conditions only. Threshold and slope are shown for each individual noise level in the visual display.

supporting **H2**. This shows that participants relied more heavily on the audio than on the noisy visual information. The opposite was observed for the near

range; the weight dropped from noise level 1 to noise level 2 and then remained constant. With weights under 0.5 in the near range *INCON* conditions, it is assumed that participants still relied on the visual information even though the display was heavily degraded, but this calls for further research and investigation.

Figure 5-4 shows the results of correlations between prior experience playing computer games, prior experience with virtual reality head mounted displays, slope and threshold of psychometric functions, and mean accuracy across all the AV *INCON* conditions. The AV1, AV2 and AV3 rows represent the participants'

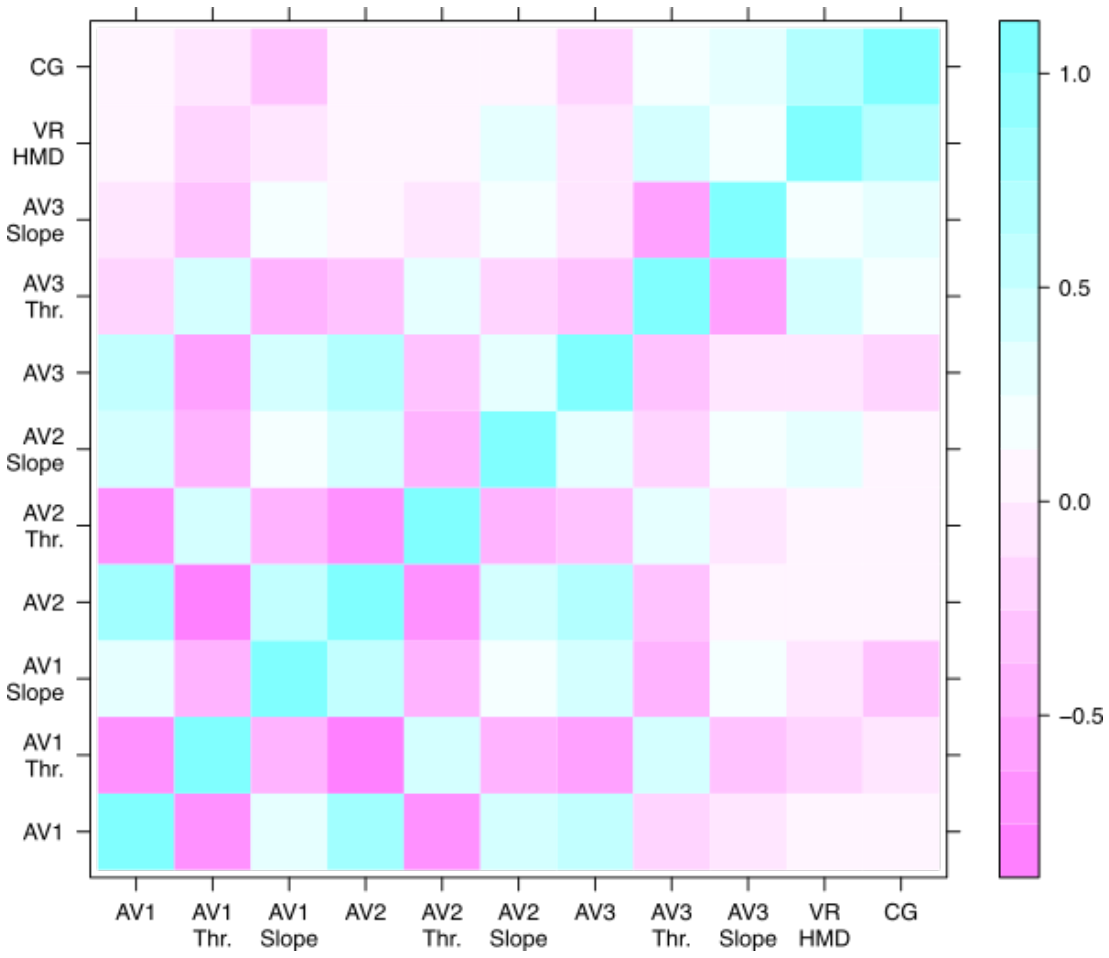

Figure 5-4: Pearson correlation matrix between mean accuracy in the experimental task, the slope and threshold of AV conditions, prior experience with a VR HMD, and prior experience playing computer games. The low correlations between accuracy and HMD usage, and between accuracy and game play experience, are indicators that our method is unrelated to either factor.

accuracy in the audiovisual conditions, with CG (computer games experience) and VR HMD (experience with virtual reality headset displays). The correlation is low ($r < 0.5$), implying that the results of our experiment generalize across the population rather than being skewed to a subset who frequently play computer games or who are familiar with VR head mounted displays. There is no evidence to suggest that the method relies on mastery of computer games, and it is independent of prior experience with head mounted displays.

5.2.4 Discussion

This study applied psychophysical analysis to explore how humans compress distance in virtual environments using HMD technology. Psychometric functions vary in two main characteristics: their slope and their 50% correctness threshold (chance level). As can be seen from Figure 5-3 and Table 5.1, column 3, the slopes of the psychometric function have fallen comparing the far range to the near range *CON* conditions (Panels A & B), and similarly in the *INCON* conditions (Panels C & D). A lower slope indicates a less restrictive dynamic range; the data vary more between the threshold value and the point of subjective equality (PSE) [196]. Participants were less accurate in their estimates (lower slope) for far range than for near range *CON* conditions, which is expected given that distance estimates are less reliable further away due to compression. However, comparing the ranges across congruence and incongruence, the data show higher slopes (Table 5.1, column 3) in the *INCON* conditions compared to the *CON* conditions. This increase in slope means participants were *more* accurate in their estimates when presented with incongruent stimuli.

In the *CON* Noise Level 1 condition, the slope of the function is 0.84, while in the *INCON* Noise Level 1 condition it is 0.95 (Table 5.1, column 3). An increased slope was observed excluding the near Noise Level 3 conditions, and the far Noise Level 1 condition. Participants were more accurate in the *INCON* conditions compared to the *CON* conditions. The increase in psychometric function slope observed between identical visual noise conditions across *CON* and *INCON* conditions means participants were more accurate in the *INCON* conditions compared to the *CON* conditions, thus supporting **H1**.

The results in Table 5.1 indicate that the weights for the audio modality were much higher in the far range than in the near range (Table 5.1, column 5). The near range audio weights are all under 0.5, meaning that the audio modality estimate accounted for less than 50% of the total distance estimate. In the far range, the audio modality estimate rose as the visual noise increased, to above 90%, meaning that participants biased their estimates to the audio modality much more heavily than the visual modality. Hence, **H2** is supported only in the far range.

Threshold values decreased in the *INCON* noise levels 2 & 3 conditions compared to their corresponding *CON* conditions (Table 5.1, column 2). Participants could better discriminate between positions in trials where the audio and visual stimuli were incongruent, however, this effect is observed only for the far range. This finding implies that incongruent presentation did not reduce distance estimate errors when the targets were close to the observer (3 $\pm$2.5 meters).

Distance overestimation, where objects are perceived as further away than they actually are, occurs in both the visual and auditory domains when targets are presented close to the observer. The crossover point, that is the point at which the observer typically moves from underestimating to overestimating and vice versa, is influenced by the constant parameters k and α fitted to the positioning function [139]. This crossover point is closely related to the specific distance tendency (SDT), the point where targets are perceived when the observer is given minimal distance cues. Anderson & Zahorik found the crossover point to be 3.23 meters in an audio-only condition, and our function is based on their parameters [139]. They do not report a combined audiovisual crossover point. The findings here suggest that the audiovisual crossover point is at a similar distance.

The compensation function was designed to compensate for distance compression by presenting audiovisual targets incongruently, specifically by using the visual component of a target as an anchor and systematically positioning the audio component of the target further away from the observer. It was applied regardless of the egocentric distance between the observer and target. In the experiment, near range trials had a reference point of 3 meters. If the crossover point is indeed $\approx$ 3.23 meters, then one may infer that in the near range condition our function

had the effect of worsening distance overestimation. As the thresholds are shifted to the right in the near range (and far range Noise Level 1 & 2 conditions), it is plausible that the function has adversely affected distance estimates under these conditions. If the crossover point can be reliably determined, the function could be modified and, instead of simply reducing distance compression by pushing the audio back (as we have shown to be effective in the far range), it could adapt to the range and reduce distance expansion in the near range by pulling the audio in front of the target's visual component.

More research is needed to investigate potential negative effects that might be introduced by using incongruence as a design tool in VEs. For example, if there is any interaction between egocentric and exocentric distance perception, it might be affected by incongruence in one domain or the other. If incongruence leads to reduced egocentric distance compression, it is unknown what effect (if any), this may have on our ability to internalize spatial maps of a scene. There is also evidence to suggest that visual experience affects internal spatial representation [162]. Further research is needed to investigate how manipulation of the audiovisual signals (in this case, intentional incongruence) may affect this internalization mechanism. Further research could also investigate the effects, if any, of incongruence on commonly reported issues with VEs such as motion sickness and sense of presence.

All stimuli were constrained to the frontal view. There is evidence to suggest that localization accuracy varies as the angle between the observer's direction and the source of the sound shifts away from the 0° azimuth. Through post-hoc analysis of a real world experiment, Chan et al. provide evidence of higher accuracy in localizing a multi-modal stimulus in both visual-only and audiovisual incongruent conditions [138]. Zahorik has demonstrated for auditory distance perception that the weights applied to various distance cues changed substantially for various positions from the frontal plane [134]. Further research is needed to investigate whether our results hold in audiovisual virtual environments where targets appear at various positions around the observer.

Basing the incongruent method on ML to compute the weights of the auditory and visual signals meant that it was not possible to include a noise free audio-

visual condition in our analysis. Future research might apply other methods to investigate different audiovisual conditions, however, there cannot be a noise free audiovisual condition in an absolute sense. The quality of the visual signal is relative and will vary depending on, for example, which HMD is used.

This study immersed participants in a sparse, minimally populated VE within a laboratory setting. It remains to be shown whether such results can be replicated in more realistic settings which could include a variety of visual cues, auditory cues and audiovisual targets.

5.2.5 Conclusion

The findings from Experiment I suggest that intentionally rendering the auditory and visual components of objects incongruently to one another could improve distance perception in a virtual environment. Having derived and tested a method for positioning the auditory and visual elements of an audiovisual target incongruently to each other, results show that the method was successful when the target was at longer range. Participants were more accurate in a distance discrimination task when the auditory and visual components of the targets were incongruent.

At closer range, where distance *expansion* may have affected the observers' perception, the positioning function actually seems to have made the distance estimates worse, which corroborates previous work suggesting a crossover point at ≈ 3.23 meters. If this crossover point can be confirmed for audiovisual targets in VEs, we can refine our method such that it is range adaptive, i.e. adapting for distance expansion up to the crossover point and for distance compression beyond the crossover point.

While the results from Experiment I are compelling, they are limited by the constrained, abstract environment used. Experiment II aimed to tackle these limitations by exploring the perception of distance in a more ecologically valid environment, and with a different measurement protocol. Before detailing Experiment II, the next section details other attempts at solving distance compression.

5.3 Solving the Distance Compression Problem

In the visual domain, manipulating the geometry of the scene can reduce the level of distance compression in blind walking tasks [10]. The technique, known as minification, involves shrinking the image to be displayed, and then rendering the resulting image so that the complete field of view is scaled appropriately. Minification affects the angle of declination (see Chapter 4, Section 4.2.4) by increasing or decreasing the angle of declination, with respect to a scaling factor (as shown in Equation 5.3).

$$\theta_{new} = \arctan\left(m \cdot \tan\theta_{orig}\right) \tag{5.3}$$

Kuhl and colleagues conducted between subjects experiments to investigate the affect of minification on distance perception, and found that participants subjected to the minification rendering technique made significantly less estimation errors than a control group (who viewed the un-minified scene). However, it is important to note that all participants still performed less than ideal estimation, demonstrating that compression still occurred regardless of minification.

There is evidence to suggest that spatial perception in virtual environments is an adaptive process. Waller et al. showed that humans can adapt to the spatial cues of a virtual environment rendered in a HMD after a short period of interaction inside the environment [144]. However, upon leaving the virtual environment, participants then experienced a phase or re-adaptation: distance estimates made in the real world after virtual environment exploration was compressed similarly to the real world. In an ideal VR scenario, this re-adaptation phase would be minimised or eradicated completely, enabling seamless transition between virtual, physical, and mixed environments. Other researchers have explored ways in which to manipulate perception of 3D space in VR. Bruder and colleagues experiment with redirected walking, a technique to leading a user immersed in a virtual world, along a specific trajectory that is not mapped directly to their trajectory in the real world. This enables virtual world designers to expand the exploration space in the virtual world to a greater area than is physically available. In tests of the application of redirected walking, Steinicke et al. found that humans can

be redirected through rotation gains applied to the camera rotation that are not 1-1 with the physical rotation of the user's head [96]. Most interestingly, is the finding that user's can be made to walk in a circle of radius greater than 22 meters yet feel as if they have maintained a straight line. This effect has been reproduced in [197].

In implementing their redirected walking technique, Steinicke et al. found that distances can also be modified in the virtual environment so that they don't match the physical world, yet are perceptually veridical to the user [96]. Experiment I in Section 5.2 demonstrated how incongruity is an effective technique for reducing distance compression. By intentionally misaligning the acoustic source from a visual stimulus (i.e. seeing and hearing a loudspeaker), participants were more accurate in their estimates of distance based in a 2 alternative forced choice (2AFC) task[3]. One shortcoming of this study was that the stimuli were always positioned directly in front of the observer (azimuth 0°). Given the evidence from prior studies showing difference in compression with respect to angular position [11], it is necessary to investigate how incongruent compensation interacts with the angular position of an audiovisual target. While incongruence is an intriguing hypothesis to the solution of distance compression, it requires further study. Studies such as that by [185] demonstrate a negative interaction effect in multi-modal virtual environments, failing to reproduce the claim of distance compression within audiovisual environments. However, it is important to note that manipulating the environment was not an independent variable of their experimental design, therefore participants' perception of distance may have already adapted to the virtual environment and the response method applied in the experiment. They do note an asymmetry in the accuracy of room size perception however. A similar asymmetry, this time between the positions of auditory and visual components of a stimulus, is the crux behind the incongruence method derived in Equation 5.2 and in the experiment reported here.

[3]See [198, 199] for details on nAFC tasks.

5.4 Experiment II: Examining Ecological Validity

Before discussing Experiment II, a few points to note:

1. Distance perception within virtual environments has been found to be compressed with respect to the real world across various studies.

2. Due to the asynchronous processing of multi-sensory information in the brain, distance compression may be reduced by systematically constructing audiovisual virtual environments with the goal of relieving this compression through misalignment.

3. Consequently, distance perception is seen as a cross-modal cognitive task, and therefore should be studied from the perspective of multi-sensory integration and psychophysics.

4. It is unclear whether incongruity is only effective in *some* virtual environments, consisting of abstract and unfamiliar stimuli, or whether the positive impact of incongruence is generalizable and applicable to more natural environments.

The first point has been demonstrated across multiple studies as discussed in the preceding sections. The second and third points have been addressed previously in Section 5.2 where a novel method of compensating for distance compression through the design of an incongruent display was proposed. In order to address the last point, I designed and conducted an experiment which looked at distance perception in virtual audiovisual environments, and analysed the data in order to determine the effect applying incongruence to the environment might have. This study had the following hypotheses:

H1: Distance perception would be compressed, causing participants to incorrectly match/adjust the location of the audio stimulus to the visual target stimulus.

H2: Distance compression would be impacted by the angle between the observer's front view and the visual target, with compression being heightened as the angle becomes smaller (i.e. parabolic or quadratic fit).

The experiment involved 40 participants (12 Female), a mixture of undergraduate and postgraduate students at a local university, with a mean age of 26 years, participated in an alignment task experiment. Spatial audio was implemented using the same custom plug-in written for the Unity Game Engine used in Experiment I. All participants were unaware of the experiments hypotheses, and had no known hearing impairments. The Oculus Rift headset was calibrated based on the IPD of the individual participant using the configuration utility from the Oculus Rift SDK. Participants were seated in a chair, and head movement was restricted using a chin rest. The hardware used to execute the software implementation was the same as the experiment described in Section 5.2.

5.4.1 Environment and Stimuli

Inside the virtual environment, participants saw a series of scenes each depicting a row of cars parked, viewed from the side. In each scene, the closest car was recorded as being 5 meters from the camera's position, measured using a tape measure. Each scene was rendered from stereo photographs captured with a Canon EOS 60D, with a 50mm lens, mounted on a custom camera rig consisting of a tripod, sliding dolly, and wooden frame (See Figure 5-5). The photographs were captured by taking a single photograph, namely the photograph for the right eye, then shifting the camera 2.5 inches to the left and taking a second photograph. 2.5 inches was chosen as it is a measured estimate for the IPD across the general population [200]. A sample image pair from our asset set is shown in Figure 5-6. Scenes 1 & 3 contained 6 cars while scene 2 contained 5 cars for a total of 17 distances across all scenes. The audio stimulus was a synthesized car horn (1 second, mono channel) that participants moved around the scene. All scenes were comparable in terms of visual cues, and the horn sound was identical in all trials.

5.4.2 Procedure

Participants wore the headset, rested on a chin rest throughout the experiment. Before beginning the experiment, each participant was guided through a tutorial

that demonstrated the input mechanism and the task for the experiment. The UI shown in Figure 5-7 was displayed throughout the entire experiment to remind participants. After the tutorial, the main experiment began. In each trial, participants would see the stereoscopic scene, and they were asked to place the

Figure 5-5: The camera rig used to take the stereo photographs.

Figure 5-6: An example camera image pair used in the experiment. The left image is the left eye view, the right image is the right eye view. Both images were rendered stereoscopically. There were three such image pairs; one pair for each scene.

Figure 5-7: A screenshot of the view from within the Oculus Rift DK2. The UI was displayed in all trials as a reminder of the control scheme. It only disappeared from view during each trial interval, after the participant had input their response and before they had begun the next trial at which point it reappeared.

sound of a car horn to match the position of a target car. To indicate the target car, a textual prompt was shown before each trial. After reading the prompt, participants would press a button on the game controller, upon which the screen would fade to black and the trial scene would fade in. Using the game controller, participants could move the current location of the car horn in a semi-circle arc in front of them, from 0°- 180°. No online auditory feedback was provided; at any time, participants could poll the audio in order to hear the horn from its current position. They could also hear the horn from a reference position: this position was 1 meter, directly in front of them. Once they were happy that the sound of the horn matched the perceived location of the target car, they confirmed their response using the game controller and moved on to the next trial. Each participant completed 30 such trials, presented in randomised order. Each trial was a combination of a scene and a target car.

5.4.3 Results

All data were processed using the R Language and Environment for Statistical Processing [195]. Plots were produced using ggplot, a graphical statistics package in R [94]. Figure 5-8 shows the results of a regression model, with output of the incongruent function plotted against the perceived distance. The raw data for each set distance is also plotted on a continuous scale. Data is averaged over all 40 participants, with trial repetitions included, governing 20 trials on average. A statistically significant effect was observed, with the model explaining over 91% of the variance ($R^2 = 0.91$, $p < 0.01$). Figure 5-9 is the residuals plot demonstrating homogeneity of variance.

Table 5.2 shows the error margins for each distance in meters over all 3 scenes. Distances that resulted in outliers were removed, reducing the number of distance targets in scene 2. Each scene was taken from a different angular perspective, with cars of different shapes, sizes, and colour. The response values follow a linear rise, with variations most notable in Scene 2. During debrief, some participants divulged different strategies for matching the horn sound to the target car. These strategies consisted of 'always picking the front of the car' to 'matching the nearest part visible' to 'matching the horn to the perceived centre of the car'.

The second hypothesis of the experiment was that distance compression would be influenced by the angular position of the visual target. In the geometry of the scene, the cars moved up in parallel, therefore their distance from the observer co-varied with the angle from the edge of the view. Thus, a repeated-measures ANCOVA was also conducted on the data for each scene, with distance as a co-variate, to analyse the effect of angular position on the response from each observer. Results are plotted in Figures 5-10, 5-11, & 5-12. The log function was applied to the response values before analysing for ANCOVA to fit the assumption of normality. Calculating means for the angle as a factor, with bootstrapped confidence intervals, resulted in a significant effect of IV 'Angle' on DV 'Response' ($F(5, 120) = 20.48, p < 0.01$) for the first scene. For the second scene, using similarly bootstrapped confidence intervals resulted in a significant effect ($F(4, 75) = 7.59, p < 0.01$). Finally, for the third scene, using the same bootstrapped method, resulted in a significant effect ($F(5, 92) = 15.92, p < 0.01$).

Distance to target (meters)	Mean Response (meters)	Error (meters)	Scene
5.00	0.34	4.66	Scene 1, 2, & 3
7.18	0.50	6.67	Scene 1
7.31	0.283	7.03	Scene 3
7.76	0.45	7.31	Scene 2
9.47	0.68	8.79	Scene 1
9.71	1.00	8.71	Scene 3
10.45	0.61	10.04	Scene 2
11.80	1.06	10.74	Scene 1
12.15	0.90	11.25	Scene 3
13.59	1.16	12.43	Scene 2
14.16	1.11	13.05	Scene 1
14.61	1.34	13.27	Scene 3
16.52	1.572	14.95	Scene 1
17.08	1.62	15.46	Scene 3

Table 5.2: Table of error margins for mean responses over each distance in the trial set. The error (distance to target - response) increases linearly as the actual target distance increases.

Post hoc tests using Tukey Honest Significant Difference testing showed significant differences in means across all pairwise-comparisons of grouped means on angle. For scene 1, Tukey HSD values for the angle of the first car paired with the third car (16.3°and 31.4°) were significant at the $p < 0.05$ level. Similar results were found for the first car paired with the forth, fifth, and sixth cars. This trend continued for the second car with the forth, fifth, and sixth cars, for the third car with the fifth and sixth, and for the forth car with the sixth. Table 5.3 shows the significance results with adjusted p values of the Tukey HSD tests for scenes 1 to 3.

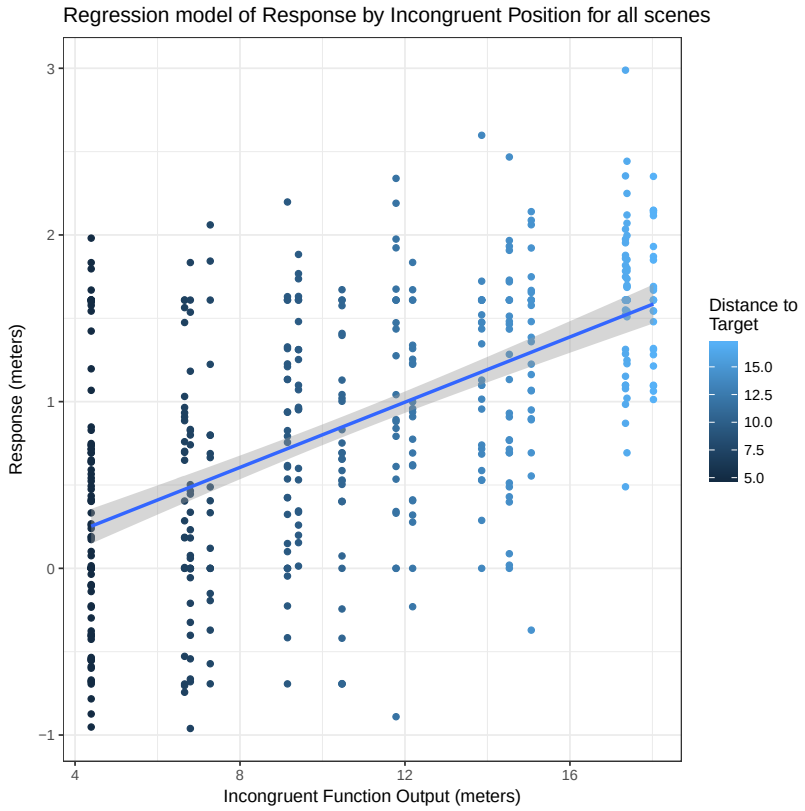

Figure 5-8: Log-Linear plot of incongruent function output against response values from participants. Real world distances to targets are shaded on a continuous scale. 0 (zero) represents the participants location. The difference in the axes represents the compression rate of participants over all scenes, demonstrating the need for calibration of the incongruent function.

5.4.4 Discussion

The aim of this study was to investigate whether incongruence would be effective in photo-realistic environments that provide distance cues to the observer. Finnegan et al. hypothesized that audiovisual distance perception could be manipulated by the application of an incongruence function that would result in the auditory and visual stimuli of an audiovisual target being rendered from different Z locations [140]. While that study was limited to a 2AFC paradigm, our current study provides further evidence in the form of a perceptual matching task. The results from the log-linear regression model on response by incongruence function show a strong correlation of computed versus measured distance perception, but

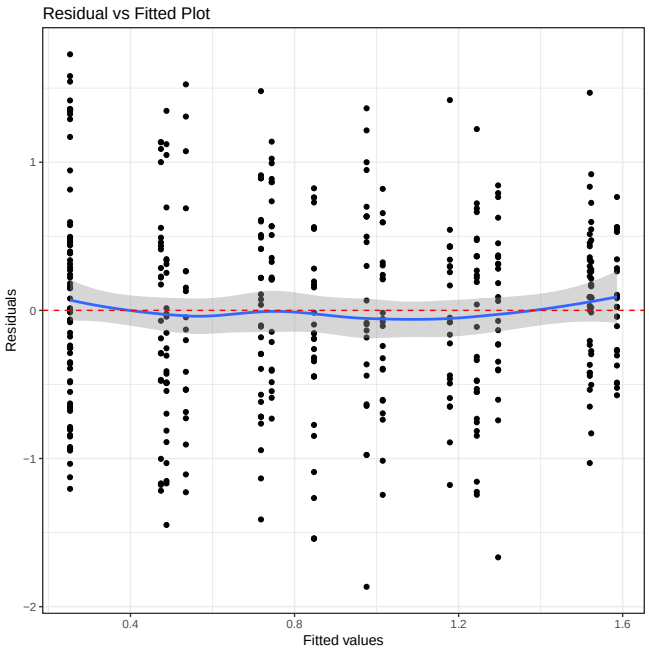

Figure 5-9: Residual plot of the incongruent function model from Figure 5-8 demonstrating similar variance across distances.

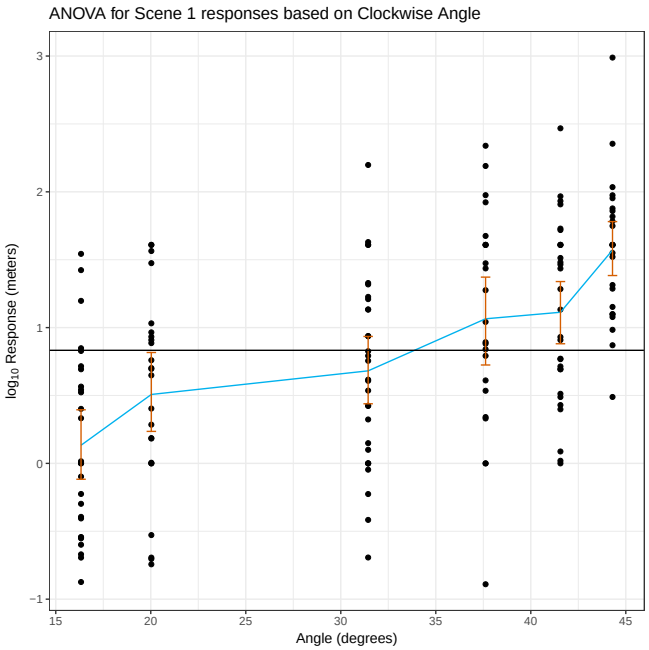

Figure 5-10: Results of ANCOVA on scene 1 between means modelled by angle from the observer, controlling for distance. Responses followed a non-linear rising trend with respect to angle, implying a quadratic fit with respect to angle.

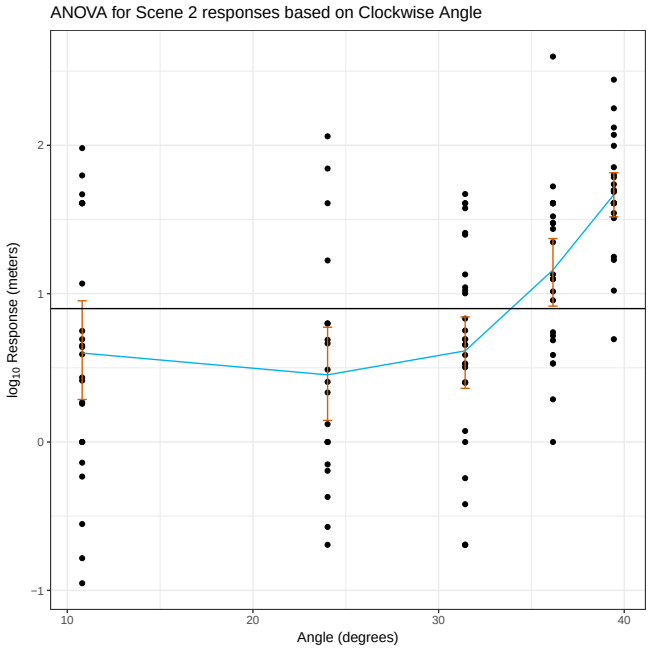

Figure 5-11: Results of ANCOVA on scene 2 between means modelled by angle from the observer, controlling for distance. Responses followed a non-linear rising trend with respect to angle, implying a quadratic fit with respect to angle.

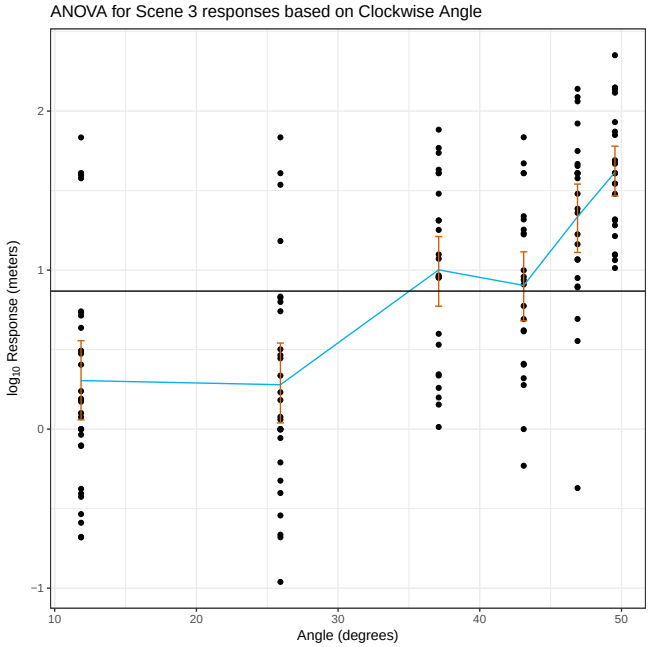

Figure 5-12: Results of ANCOVA on scene 3 between means modelled by angle from the observer, controlling for distance. Responses followed a non-linear rising trend with respect to angle, implying a quadratic fit with respect to angle.

Scene 1		
Angle A	Angle B	adj p
16.33	31.44	p <0.05
16.33	37.62	p <0.05
16.33	44.30	p <0.05
20.03	37.62	p <0.05
20.03	41.56	p <0.05
20.03	44.30	p <0.05
31.44	44.30	p <0.05
37.62	44.30	p <0.05
41.56	44.30	p <0.05

Scene 2		
Angle A	Angle B	adj p
10.80	36.16	p <0.05
10.80	39.44	p <0.05
24.02	36.16	p <0.05
24.02	39.44	p <0.05
31.43	36.16	p <0.05

Scene 3		
Angle A	Angle B	adj p
11.86	37.10	p <0.05
11.86	43.10	p <0.05
11.86	46.90	p <0.05
11.86	49.54	p <0.05
25.93	37.10	p <0.05
25.93	43.10	p <0.05
25.93	46.90	p <0.05
25.93	49.54	p <0.05
43.10	49.54	p <0.05
46.90	49.54	p <0.05

Table 5.3: Tukey HSD adjusted p values for pairwise comparisons across all angles of scene 1. Angle A and Angle B represent angles extended between the participants' front facing orientation and cars in each scene. Only statistically significant pairs are shown.

the misalignment of the axes in Figure 5-8 shows the need for a calibration mechanism. Although participants perceived depth in the scene, they still perceived severe compression. As the incongruence method is based upon fitted curves of perceptual studies, the α and k parameters are decided based on past data. This implies that, as different contexts seem to produce varying compression results [8, 144, 160], the incongruence function itself is context sensitive. Therefore, an adaptation, or more appropriate, a 'training' period is required to calibrate the incongruence function to the observer's individual rate of compression. The perceptual matching task used here could be adapted into an appropriate calibration procedure. More research gauging the flexibility of incongruence in other environments will lead to a more complex, general function.

The fact that compression occurred is expected, as numerous past trials attest to this [139, 153]. Severe distance compression has been observed in at least one previous study which also used a head mounted display [148]. It is unclear why such severe compression occurred here, but one reason may be that observers had no reference distance in the visual modality. The only reference given was in the auditory domain consisting of the horn sound positioned 1 meter in front of them. Perhaps participants saw this as a clue to the actual location of the cars, and

thus used this as a basis for generating their estimates. Another reason the cars were perceived as being much closer could be due to the binocular disparity kept constant throughout the experiment and across all scenes. Results from Turner et al. suggest that 3D audio may be used to enhance the perception of binocular depth [192]. Combining this evidence with the use of a constant IPD value in the presentation of stimuli in our study, is it viable to suggest that there are subtle cross-modal interaction effects between vision with respect to IPD and binaural audio. However, more research, specifically in the form of a controlled experiment specifically looking at the interaction between IPD and spatial audio position on depth perception in binocular environments, would be needed to confirm this.

As noted, some participants chose different strategies for positioning the car horn. Trials where participants positioned the sound to the front of the car would result in different distance response values compared to participants who chose to just put the sound to the centre or rear of the car. Thus asking participants which part of the target object to match to affects the results. Specifying the exact part of the target object to match would lead to reduced error in the calibration phase.

The second hypothesis of Experiment II was that angular position of the target would affect the compression rate was also demonstrated in our results. Controlling for distance, angle had a positive effect on the compression rate, increasing the distance response from observers as angle increased. This implies that distance compression gradually declined with increasing angle. Results from studies in manipulating source distance perception in binaural auditory listening tests also show main effects of angular position [11], and our results are in agreement. Note that the adjusted p-values of the Tukey test results seem to follow a pattern; statistically significant results were observed as the angle increased, but when the *difference* in pairwise comparisons was greater than a noticeable threshold. This seems to suggest that there is a kind of 'just noticeable difference' (jnd) in compression based on angle. Jnds are known in auditory localization studies (usually referred to in these studies as minimum-audible-angles) as well as discrimination studies [9, 16, 87, 165]. However, there is not enough statistical power in this study to confirm this with a post-hoc test. Again, a controlled experiment specifically testing is required to confirm the hypothesis.

One theory we suggest based on the angular effect on compression data is that distance perception in VR follows a parabolic curve with respect to azimuth angle. Theories which attempt to explain distance compression note the effect of angle of declination on perceived distance [10, 161], but to our knowledge no study has specifically proposed a *systematic* effect of angle on distance compression. Contrast this with the fact that the incongruence function was indeed a *good* predictor of perceived auditory location; this seems to contend the notion of varying compression with respect to angle. Of course, while it correctly predicts the perceived distance, it did not correct for the compression observed across all scenes. The parabolic curve theory would suggest that the incongruence function should be adapted, namely by introducing another parameter into the function for the azimuth angle subtended from the observer to the target. Gilinsky gathered data in a real world setting that best fit a hyperbolic curve of distance perception [201]. Loomis et. al note that while the study has been replicated, distance perception has followed a more linear curve when different measurement protocols have been used [6]. It is interesting to note then how the parabolic theory may apply in VR but not in the real world with particular measurement protocols. More research is needed to elaborate on this further, and to replicate these findings in other virtual environments (e.g. non HMD based environments).

5.4.5 Conclusion

Previous studies in audiovisual distance perception in VR report compression. This study explores distance perception with the goal of *reducing* compression through cross-modal compensation based on a theory that distance compression may be compensated through intentionally misaligning the audio and visual stimuli components of an audiovisual target.

We tested the hypothesis that distance perception can be predicted by an inverted power function whose parameters were based on prior studies in distance compression. We replicated the finding of distance compression measured using a perceptual matching task. Our results showed a log-linear fit for perceived location for a given distance, suggesting that compression is a *systematic* perceptual phenomenon whose parameters may be determined. Determining these parame-

ters has beneficial impact to VR practitioners and designers of VR environments, by recommending a mapping between the acoustic and visual spaces that result in close to real world distance spatial perception.

The results suggest that distance perception, measured in an analogue alignment task, is accurately predicted by the incongruity function introduced in Section 5.1.2. However, this function does not take 2D spatial location (i.e polar coordinates) into account. We also found a statistically significant effect of target angle on distance compression. This compression seemed to follow a parabolic curve, suggesting that compression may be compensated further with a modified version of the incongruence function. In the next chapter, we'll look at incongruity in the context of mobile environments, where the factors of self motion and proprioception interact with our incongruent function.

Chapter 6

Distance Compression in Mobile Reverberant Environments

"When we try to understand reality, we experience some of the Limitless Giving–it blows our mind–and then we come up with some way to receive it, integrate it, and live with it–which is Drawing the Line."

Eric Kaplan

As we saw in Chapter 5, distance compression in audiovisual virtual environments can be compensated by exploiting the mechanism through which human perception integrates cues. However, one important point to note is that the environments I've addressed thus far have been *static*: the observer was sat in a chair and the environment itself was minimal, with no moving parts (i.e. no dynamic objects). While certainly useful, it is difficult to extrapolate the effect to more dynamic environments representative of the kinds of applications VR will be applied in. For example, an architect may wish to visualize her model building and walk around the viewing. Her perception of the scene will thus be in a closed-loop feedback system, where the visual field will update as she walks around. Also, some rooms may have varying acoustic reflectance models,

meaning as our architect explores the 3D model, she would perceive transitions in both the visual and the acoustic virtual space.

Of course, our architect may not be the only agent in a virtual environment. A virtual meeting room, may have other people digitally represented, who may rise from their chair and walk around the room, or towards the window (with a clicker to change the outside environment of course). As they move, their relative position with respect to some observer's fixed point of view will change, affecting the observer's perceived distance and general spatial perception of the scene they are in.

In this chapter, I focus my attention on dynamic virtual environments which better represent the intended application domain of VR systems. I discuss distance compression with respect to studies that involved dynamic environments, typically where the observer moved around either by use of a joystick/joypad or actually walking on the spot or in real space. I then detail my final study, which sought to establish the interaction between proprioception, audition, and vision in the context of dry and reverberant environments.

6.1 Proprioception in Virtual Environments

In order to maintain immersion and situational awareness, the system should update the visual and audio environments as the observer moves around. Traditionally, such an update is enabled through the application of a game pad, joystick, or other peripheral. With respect to distance perception, studies using peripherals have also demonstrated compression issues. For example, Murgia & Sharkey found distance to be compressed in a CAVE system when participants were tasked with moving a virtual sphere to the same location of a previously rendered virtual target [122]. In an experiment investigating depth perception in audio using a head mounted augmented reality device, distance was also found to be compressed with the use of a both a wand and a joystick for input [190].

As mentioned in Chapter 4, Section 4.2.3, distance perception has been shown to interact with the measurement protocol used. Various measurement protocols have been applied in studies over the past few decades; some of which involve

physical body motion and some of which do not. Studies using indirect input mechanisms such as peripheral controllers typically use protocols like verbal response, perceptual matching tasks, and alignment tasks. Such protocols require the participant to at most move a single limb in order to control a virtual object as an implicit measurement of their response to a distance perception trial. Other studies have employed techniques such as blind walking, cycling, and wheelchair input.

The main differentiating factor between the former and the latter examples is the addition of motion in the latter cases. By having participants physically move their bodies, studies have been able to further specify factors influencing distance perception. Work by Sun et al. observed participants engaging in a distance perception task involving proprioceptive cues [191]. Their study involved cycling on a fixed-position bicycle while wearing a head mounted display. The bicycle was equipped with an infra-red sensor for capturing pedalling speed, and the HMD displayed the same image to both eyes, with a FOV of 60°. By taking the sensor information, Sun et. al modelled the relationship between visual and proprioceptive cues through a gain factor applied to the optical flow. As participants cycled, the optical flow was systematically manipulated by applying a gain factor to the speed of the visual environment update (termed optic flow gain or OFG). This provided a method to explore the impact of proprioception and vision upon distance perception independent of each other and when combined.

In their experiment, Sun et. al manipulated the OFG in order to assess how participants judged the distance pedalled: whether visual cues or proprioceptive cues were dominant. The study design represented 4 conditions: 2 congruent conditions, 1 with pedal and vision and the other with vision and mouse input. There was a single incongruent condition, where the OFG was applied to present trials where participants experienced both $1 \leftrightarrow 1$ mapping of visual update with pedal speed, and a $V = kP$, where k was chosen at random. The final condition turned off the HMD display, so that only proprioceptive information was available to the participant. These 4 conditions allowed a baseline of performance to be established, and then to investigate interaction effects between vision and proprioception. The results showed interaction effects between the OFG and the path length in both the incongruent condition and the mouse and vision

condition. More so, in the non-visual condition, the error rate was comparable to the congruent condition; this implies that proprioception alone was enough to result in accurate path length estimation.

A later study by Campos, Butler, and Bülthoff examined multi-sensory integration[1] (MSI) in the context of walked distances [154]. Rather than taking a random approach to the incongruent environment parameters, this study employs arbitrary fixed gain values of 0.7 and 1.4 in the optic-flow/proprioception model. Four similar conditions to the study from Sun et al. were used: Congruent, Incongruent, Body only, and vision only [191]. The results indicated differences between the congruent and body only conditions, and the congruent and visual only conditions. Interestingly, with respect to incongruent conditions, in a 3 x 4 ANOVA between gain factor and distance, there was no interaction effect between the two. This implies that the variance in the participants judged distances were the same across gain factors and distances travelled. In their discussion, Campos and colleagues point out that combined estimates (vision and proprioception) did approximate the body only estimates rather than the visual only estimates. This corroborates the results from Sun et al. where proprioception seems to be a *heavier weighted* cue than vision with respect to motion in virtual reality systems [191].

6.1.1 Dynamic Effects in Walking Experiments

The studies discussed above took place in environments where participants walked in straight lines. There is also evidence regarding cross-modal perception when participants do not walk in straight lines, such as work by Frissen et al. with respect to the integration of proprioception and vestibular cues; cues that relate to our sense of direction and orientation [202]. In their study, participants were tasked with continuously pointing to a target while in motion. The apparatus consisted of a clockwise rotating treadmill which can either operate automatically or be pushed by a participant. The treadmill is consists of two moving components; the floor and a rotating handlebar for pushing. Their experiment studied vestibular cues and motion cues similarly to those above; independently

[1] See Chapter 3, Section 3.3.3.

and in tandem. In a vestibular only condition, the treadmill operated automatically: thus participants were in a passive motion state, perceiving dynamic cues from their vestibular system while standing in place. In a motion only condition, only the floor moved with respect to the participants motion, like a standard unidimensional treadmill. Finally, in a combined condition the participant operated the treadmill via handlebar, and also subject to a congruent/incongruent design. Incongruent conditions used the same gain factors as Campos et al. [154]. ANOVA results indicated a main effect of gain on the pointing rate. With respect to integration of proprioception and vestibular cues as stated in the maximum likelihood estimation theory proposed by Ernst & Banks, vestibular cues were weighted more than proprioception [109]. However, it should be pointed out that this study took place in a real world, with a real object in the physical room as opposed to the participant wearing a HMD.

A treadmill imposes restrictions on the area participants can walk. For virtual environments, it is more appropriate to allow participants free reign to wander unhindered. This can be achieved through the application of motion tracking technology, in particular non-obtrusive optical tracking technology. Steinicke et al. observed participants walking in the real world while wearing HMDs, with their motion tracked via an optical tracking system [96]. In contrast with the heavier weighting of proprioception with respect to distance observed in the previous studies by Sun and Campos et al. [154, 203], Steinicke and colleagues discuss how vision dominates when proprioception and vestibular cues disagree. This implies that the interaction between visual, vestibular, and proprioception cues is more subtle, depending on the context and the motion of the observer. Steinicke and colleagues aim to exploit this unintended bias towards vision in individuals in order to nudge people to walk along a curved path which, to the observer, is perceived as a straight path. First described by Razzaque et al., this technique is known as *redirected walking* [204].

In their study, they designed a scenario where participants were tasked with exploring a virtual world while gain factors were applied to their motion. Steinicke et. al describe their model for redirected walking as used in their previous work [205]. By applying gains to the unit vectors representing the walking direction, strafe direction, and up vector, Steinicke and colleagues redirect the participant.

This model, that they term the Locomotion Triple, is used to represent the way gains are applied, with the aim that redirection is *unnoticeable*, and does not impact on the observers perception of the environment with regards sense of presence, immersion, and spatial representation. In a series of experiments, participants were asked to explore the virtual environment while the Locomotion Triple was subjected to translation and rotation gains. The system was evaluated through a 2AFC procedure with constant stimuli method[2]; after each trial, participants were tested on whether they perceived the perturbation applied to the triple or not.

The results showed detection thresholds for gain factors of 0.67 and 1.24 for turning left and right, as well as gains of 0.86 and 1.26 for straightforward movements [96]. This suggests that participants could walk 1.26 times the physical distance in a virtual world without noticing. This undetectable incongruity in the physically walked space versus the virtually walked space suggests that an incongruent environment may also be beneficial in the context of proprioception in VR. Given the evidence from navigation through audio, namely the efficacy of audio as a guiding tool [206, 207], the next study investigated the interaction between audio-visual incongruence and proprioceptive feedback in a virtual environment. With the evidence of reverberation as a strong absolute distance cue [208, 209], we also wanted to investigate the interaction it would have on the perception of distance in a virtual environment.

6.2 Experiment III: Incongruence in Dynamic Environments

Experiment III investigated the effects of reverberation and motion on distance perception, and their impact on the incongruent method derived in Chapter 5, Section 5.1.2. A simple motion tracking system was developed and deployed in a laboratory environment.

[2]See Chapter 3, Section 3.3.1.

6.2.1 Experiment Apparatus

As in the studies reported in Chapter 5, the Oculus Rift DK2 was used as the virtual reality head mounted display. For this experiment, the headset was soft-modded by fixing a Sony PlayStation Move© controller (PSMove) to the front of the headset, above the housed screen display. During each experimental trial, the PSMove would light up: a simple computer vision algorithm (detailed in Appendix A.2) tracked the position of the PSMove by reading in the stream from the Sony PlayStation Eye© camera (PSEye), and searching for the location of the PSMove. The FOV of the PSEye was kept constant to 75°, and the full width of the camera's image was mapped to a 10 meter virtual environment. Thus a complete walk across the width of the image resulted in walking 10 meters in the virtual world. Participants wore the headset at all times during the experiment, which lasted on average of 80 minutes per participant. Headphones were placed on participants before the beginning of the experiment, after the briefing, and also remained in place for the complete duration.

Two machines were used to drive the experimental software; one machine handled the tracking module using a computer vision software tool, which tracked the 3D position of a glowing object of interest in real time. In one set of conditions (namely, those where the participant would physically walk) this machine also rendered the audio based on the the participants' current position. The audio intensity would update depending on the position of the tracked object, which represented the participant's head and thus standing position throughout the experiment. This position was then sent across a local area network (LAN) to a second machine which implemented the Unity application which rendered the 3D environment. In the conditions where the participants moved by using a gamepad, the same machine was used to render the graphical environment as well as the binaural audio. However, the PSMove remained in place on the participant's head, and the network cable for the LAN was not disconnected.

Participants were split into two groups for a 3 factor, mixed design: every participant, irrespective of group, participated in 6 experimental conditions. Whether or not reverberation was applied to the sound stimulus was split across groups as the between-subjects factor. Half of the participants heard the sound stimulus

Reverberation			No Reverberation		
Environment	Gamepad	Congruency	Environment	Gamepad	Congruency
Audiovisual	Yes	Congruent	Audiovisual	Yes	Congruent
Audiovisual	Yes	Incongruent	Audiovisual	Yes	Incongruent
Audiovisual	No	Congruent	Audiovisual	No	Congruent
Audiovisual	No	Incongruent	Audiovisual	No	Incongruent
Visual	Yes	NA	Visual	Yes	NA
Visual	No	NA	Visual	No	NA

Table 6.1: Experimental Conditions in the mixed design experiment.

with Schroeder reverberation[3] applied, while the other half did not. Table 6.1 tabulates the 3 factor mixed design experiment conditions.

6.2.2 Procedure

Participants entered the lab, and were given a briefing on the experiment procedure (task, not the IV manipulation or the experiment's goal). After briefing, participants were handed the HMD and asked to place it on their head. After this, the closed back headphones were placed to cover their entire ears, ensuring the left phone was placed over the left ear, and similarly for the right ear. The presentation of conditions were randomized: if the first set of trials was the motion set, participants were then directed with the aid of the experimenter to the starting point. If the first set was the gamepad trials, they were directed to sit down in a chair in front of the machine.

In the motion trials, each trial began with the experimenter announcing the trials stopping point (TSP). After this, participants were free to listen and observer the stimuli. The moment they decided to start walking however, they were instructed not to stop, instead completing the trial in one complete forward motion. Their point at which they stopped walking (participant stopping point or PSP) was then recorded. They were also told not to move backward: if they overshot their intended stopping point, they were instructed to announce so but stay stationary regardless and not move back. In the gamepad set, the participants

[3]An artificial colourless reverberation, frequently applied in synthesizers, developed in 1960 by Manfred Schroeder [210].

were programatically restricted from moving backward as the application only responded to a forward motion with the gamepad's thumbstick. The task was identical to the motion trials. Each task set consisted of 45 trials; 15 possible speaker/stopping point combinations, each with 3 repetitions.

6.2.3 Results

In computing the compression rate, PSP was subtracted from the location of the speaker's position (SP), and this result was then subtracted from the TSP (i.e where they were asked to stop in the trial). The formula is given in Equation 6.1.

$$E = TSP - (SP - PSP) \tag{6.1}$$

This gives us the error rate with respect to each participant's perspective of a speaker position for a given trial stopping distance (i.e people stopped at 3 meters when they were asked to stop at 2 meters).

The results of three factor mixed ANOVA, taking input, congruency, and trial stopping position as within variables with reverberation as a between variable did not reveal any statistically significant effects of input modality, congruency, or the interaction between. The ANOVA showed no statistically significant main effect of reverberation on the error computed in Equation 6.1, $F(1, 14) = 0.50, p = 0.49$. No interaction with congruence or PSP was observed. In order to further analyse the effects of the reverberation on the stopping distance, the entire data was aggregated together, summarized by Participant ID and Reverb factor. A welch's unequal variances t-test was conducted on the two groups of participants, reverberation and no reverberation. There was no main effect of reverberation on PSP, $t(11.646) = 0.73, p = 0.48$.

With respect to TSP, Mauchly's test indicated that the assumption of sphericity was violated, $\chi^2(9) = 136.75, p < 0.01$, therefore degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity ($\epsilon = 0.27$). The results show a statistically significant effect of TSP, $F(4, 56) = 6.30, p < 0.01, \eta^2_G = 0.055$ and are plotted in Figure 6-1. These results suggest that participants were more accurate the further away they were asked to stop from the loudspeaker.

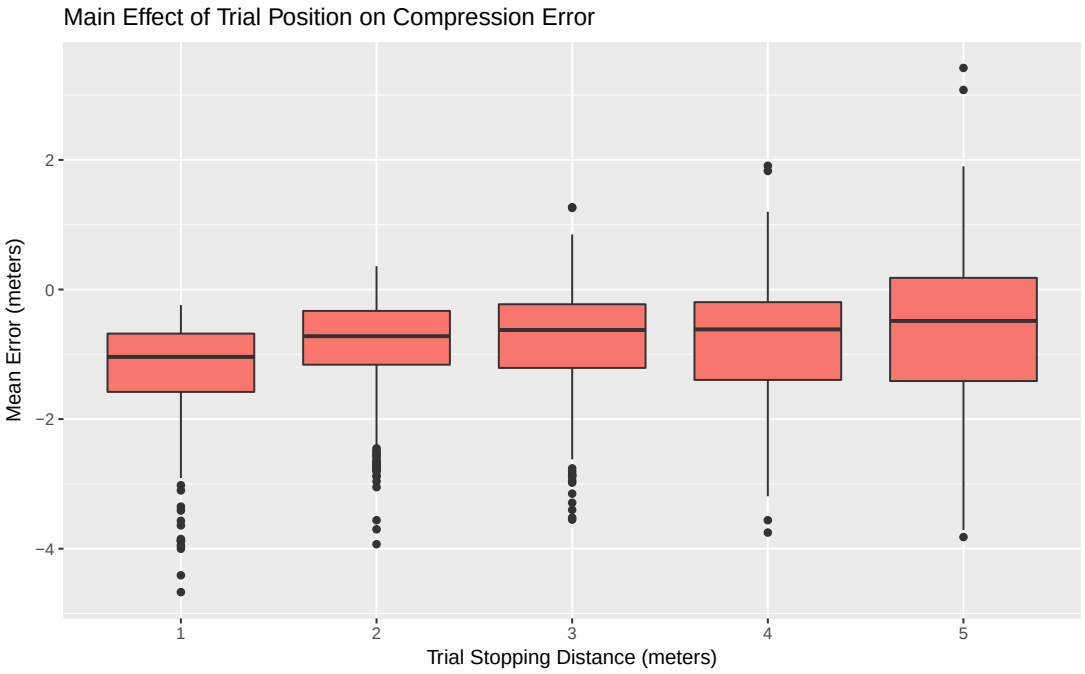

Figure 6-1: Results for ANOVA showing main effect of trial stopping distance on mean error. The trial stopping distance is the distance from the target loudspeaker participants were asked to stop at. Mean error is the average error value across participants for a given stopping distance. Negative error means participants stopped further from the target. Thus an error closer to 0 (zero) $\rightarrow$ participants stopped at the correct distance.

6.2.4 Discussion

The aim of the experiment was to investigate the impact of incongruity in immersive virtual environments when people are motion tracked, and when participant's experience reverberant, or non-reverberant auditory stimulus. In all conditions, participants moved forward the same way. For example; in walking conditions, only the dimension of motion was tracked and updated in the HMD. In controller conditions, head bopping was removed and the camera simply panned forward into the scene. In the real world, as people walk forward, their head typically sways left and right and bobs up and down. This has been implemented, as well as removed in previous studies [144, 154], in order to control for the proprioceptive feedback to the observer.

In this current study, as participants moved forward, the only difference between the motion condition and the gamepad is proprioceptive feedback perceived from leg movement and forward motion in the HMD, making it similar in design to that from Waller & Richardson [144]. Their experiment consisted of a training phase, with distance estimates taken before and after the training phase and compared across varying feedback conditions. Their results showed a slightly higher accuracy for the condition where individual bobs and sways were removed versus a body and individual bobs and sways, in the pre training phase. However, this observation reversed after the training phase, with the condition of bobs and sways removed resulting in decreased distance accuracy. Given that the only change with respect to the pre and post phases was the training phase between, Waller & Richardson concluded that the effect of training was beneficial to distance estimates through proprioception. Even though no explicit training phase was given in this experiment, there was still a statistically significant effect of motion, with proprioception affording better performance. These two results together hint at a more general effect of multi-sensory integration with respect to distance perception in virtual environments.

Reverberation was split across groups in order to reduce the length of the experiment. There was no statistically significant effect of a reverberant environment on the error rate of participants. A null result of reverberation is peculiar, as it contradicts previous findings in the literature: with respect to audio only environments, it is considered to be an absolute cue to distance and typically results in better distance perception [86, 192, 131]. In audiovisual estimates of distance in real world environments, reverberation is a little more tricky as most studies tend to use impulse response recordings from an anechoic chamber, or taking place in controlled environments with a minimal RT_{60} time [138, 153]. One clear noticeable difference is the reverberation method applied in the experiment here and in most other work. Previous work in 3D audiovisual environments tends to capture a binaural response for the environment, and then reproduce this at runtime during trials. In Experiment III, artificial reverberation was used, which does not take the physical geometrical properties of the virtual environment into account. The delay mechanism employed does not implement specular nor diffuse reflections from walls. It is difficult to assess the impact of reverberation with a

closed-feedback walking design in the context of previous work which has studied stationary participants and sound objects [208, 131, 211].

Based on the studies reported in Chapter 5 Sections 5.2 and 5.4, it was expected that incongruence would have a positive effect on reducing the compression rate of participant's while wearing the headset. However, no effect was observed in the data for the application of incongruence. There are a number of noteworthy differences between the previous two studies and the current one:

1. Experiments I and II involved static environments, with objects in the scene remaining stationary (with respect to source location) and no movement from the observer was permitted. Experiment III involved motion, both physical and through the use of a gamepad. In this sense, Experiments I and II can be categorized as ***static*** while Experiment III is ***dynamic*** in design.

2. The measurement protocol was different. Experiment I employed a psychophysical forced choice design. Experiment II took an alignment task as a distance measure. While the scene was updated, participants could only poll for the latest position of the sound source. In experiment III, a walking protocol was used which implemented a closed-loop feedback mechanism.

3. Experiment I was a purely virtual, yet abstract environment meaning there was no semantic relationship between the audio and the visual streams. Experiment II used real camera footage as visual stimuli for photorealism, but maintained a purely virtual (synthesized) audio stimulus. Experiment III was a fully audiovisual virtual environment. This made it not only dynamic, but also more specific and ecological than the previous two with respect to virtual environment design.

With these observations in mind, a number of issues arise. As stated in Chapter 5 Section 5.4, it is clear that the incongruent function could not account for angles deviating from 0°azimuth. It is reasonable to postulate that a more complex function is also required for application in dynamic scenes. Future work can address this by studying distance perception in these dynamic environments in order to develop a more general incongruence function.

One interesting and unexpected finding was the effect of TSP. A statistically significant effect was observed, with the graph in Figure 6-1 showing a clear trend towards earlier stopping distances. This is an unintuitive finding: the closer participants were asked to stop, the greater error in their stopping distance. However, note the reasonably small η^2 value for the effect size. Considering that participants always began at the same starting point, this implies that as participants walked *further*, their error with respect to stopping distance was *greater*. Moving closer to the target had a negative effect on the distance perception, meaning that participants were better at discriminating greater relative distances between themselves and a target object. Results from Mohler et al. indicated that participants in a virtual environment had shorter stride length and a slower walking velocity [8]. Perhaps that the difference in stride length and velocity would impact their perceived walking distance in our experiment, explaining the greater error as participants walked more. Future work that specifically measured these parameters using a similar measurement protocol would give deeper insight into the role of perceived and actual walking velocity while wearing a HMD.

One study reported in the literature bearing resemblance to the one reported here is that by Paquier et al. [124]. In their study, participants viewed and heard a virtual loudspeaker and had to input how far away they perceived the speaker to be. These responses were input using a keypad connected to the machine running the software, correct to one digit after the decimal point (to the centimetre). Paquier opted for this response measurement method in order to control for the directional cues that are perceived in perceptually directed action [51], as measurement method has been shown to influence distance perception [67, 159]. Paquier and colleagues found a statistically significant main effect of target distance, which is comparable to the trial stopping position effect observed in our experiment.

6.3 Conclusion

The results from Experiment III leave much food for thought. Some corroborate with past studies, namely the effect of walking versus using a gamepad, yet others are much more subtle. The null result of reverberation is peculiar, yet this may

be due to the clear separation between the audio and visual displays, combined with the simplistic approach to reverberation applied. The positive result of the trial stopping position suggests a subtle interaction between proprioceptive versus optical flow feedback, and has also been identified in the work from [8, 154, 202, 205].

This being said, the aim of Experiment III was to investigate the effect of incongruence in dynamic, virtual environments supporting the observer's motion. The results demonstrated no statistically significant main effect of incongruence on the perception of distance. Previous work suggests multiple factors that impact distance perception [2, 9, 137], and motion has been identified as a factor, as mentioned previously. Experiment I applied the incongruence function in static environments, with a stationary participant. The function itself was derived from previous research conducted in similar environments [139, 212]. Experiment II demonstrated that the function has contextual constraints. Experiment III seems to employ another set of constraints. In conclusion, it is clear that the incongruence function requires adaptation to various contexts, and requires further derivation to include more parameters from environmental factors and factors individual to the observer. Future work aiming at identifying, parameterising, and including these factors in a more complex model may result in an incongruence function derivation that can compensate for distance compression in more general virtual environments.

Chapter 7

Conclusion

"Every day is alone in itself.
Whatever enjoyment I've had,
and whatever sorrow I've had."

Henry Molaison

The aim of this chapter is to take the narrative to a full conclusion, discussing the core thesis defended in these pages regarding incongruence with respect to the studies conducted. First, I give a summary of the thesis, highlighting the journey from background review, through to sequential experimental design with concurrent software engineering commitments. I then give more detail regarding each study in the context of the research questions outlined in Chapter 1. Finally, I conclude with some limitations of the findings (no free lunch), and give some avenues for future work.

7.1 Thesis Summary

In summary, this thesis is motivated by the peculiar observation that humans reliably compress the egocentric distance towards objects in virtual environments. In studying this problem, I generated a set of research questions that aimed at addressing the compression problem by offering a solution. Using these research

questions as guides, I synthesized a method for correcting distance compression artefacts in virtual environments through an extensive, cross-disciplinary review of relevant literature in the fields of human computer interaction and the psychology of perception, most notably the domain of psychophysics. After identifying a potential solution, I then set out to design a series of experiments which tested the hypothesis of incongruent environment design. The novelty of my contribution is its context; that of audiovisual virtual environments rendered in head mounted displays. Each experiment built upon its predecessor, culminating in a thesis to defend incongruent environments as a solution to the distance compression phenomenon.

Incongruence has been specifically researched in the context of head mounted displays, yet it is contended to be generalizable. Nothing in the solution specification dictates a hard dependency to the visual or even the auditory display. While HMDs and binaural audio over headphones were the chosen displays in this thesis, I note that there is nothing expressing their necessity in the incongruent function. The incongruent function is parameterised by the position of the visual component and the audio component of an audiovisual object in a given scene. It is completely agnostic to the display type (note that the inter-pupillary distance was even kept constant in Experiment II), and in theory should generalize across other display formats and hardware. As distance compression has been observed in CAVEs as well as LSIDs: while out of scope of this thesis, it would be interesting to validate the application of incongruence to these scenarios.

Incongruence in audiovisual VR was initially investigated in Experiment I. Here, I designed an experiment grounded in psychophysics to understand the impact of incongruence on the process of multi-sensory perception, as described by the theory of multi-sensory integration. Having declared the experiments purpose, and rationalizing it with respect to solving the distance compression problem, I designed an experiment adopting a 2AFC paradigm. The experiment held a within-subjects design, testing the effect of incongruence on repeated trials from a pool of participants. Having designed the experiment, I then set out to build a system that would allow me to implement my experiment. Finding no existing software, mainly due to the cutting-edge nature of the display hardware and the relatively young age of virtual reality in the scale of large commercial use, I

developed my own system based on existing technologies. Employing this custom implementation, I conducted the experiment in various locations, both on site at my partner company and at the university. The main outcome of the experiment was that incongruence did indeed have an effect on the weighting schemes of the perceptual mechanisms of each participant. This was intriguing, yet had its own set of limitations: the environment was abstract and the 2AFC paradigm is very strict and controlled.

Experiment II aimed at further investigating the efficacy of incongruence by moving away from a 2AFC paradigm and instead focusing on an alignment task. In order to implement the experiment design, real photo assets were obtained through stereo photography of a real world location close to the university (namely, the shared staff-student car park). After the assets had been acquired, a new environment was created in which the photos could be rendered inside the headset, creating a stereoscopic perspective. Combined with a semantically congruent audio component, namely pictures of vehicles with a corresponding horn sound, each audiovisual scene was designed as close to the real world as possible, with respect to real viewing conditions. The results from the second experiment showed the predictive power of the incongruence function; corroborating previous results, I found that the incongruence function was a good predictor of the compression rate of participants. This second, photo-realistic environment, resulted in more compression compared to that in the first experiment. This led me to believe that a calibration phase should be applied to the incongruent compensating method. A second set of results indicated that distance compression is not linear with respect to angular position around the observer's field of view. As targets appeared further to the sides, away from the 0°azimuth point, compression seemed to further exaggerate. This implies that the compression function in its current form is not powerful enough to correct for compression in the periphery of an observer's view.

Experiment III moved away from the static, controller based environments of the first two studies in order to investigate the relevance of proprioception in the context of incongruent environments. Another interesting factor was that of reverberation: in the previous two studies, the auditory stimuli were always processed in dry form, as sound would propagate through an open space. This

was sufficient in the abstract environment of the first experiment, and applicable to the outdoor environment of the second experiment. However, this third experiment was designed to maintain the strong visual cues of an indoor environment, namely that of a long corridor with some reflective materials on the walls. I thus extended my software system to incorporate extra processing to the audio signal, applying simple delay filter reverberation to the sound.

In order to capture the observers own motion in the virtual environment and update their position, I also built a subsystem that integrated a simple computer vision motion tracker into my existing custom incongruence system. I then designed an experiment that could test the interactions between incongruence and proprioception, in the context of reverberant and non-reverberant environments. The results of this final study was that proprioception seemed to dominate over a standard gamepad input. No main effect of incongruence was found in the data, but an interesting effect of the stopping position was observed. This fed a stimulating discussion into what a general incongruence method might look like, as well as generating potential experiments for future study addressing this issue.

In the next few sections of this conclusion, I discuss all three studies with respect to the research questions specified in the introduction. I finally conclude with some limitations of my findings, and discuss some future work in the area of distance compression in audiovisual virtual environments.

7.2 Discussion of Findings & Contributions

The core finding of my research is a ***method for designing audiovisual virtual environments such that the phenomenon of distance compression is reduced***. While previous studies have looked in detail at the compression problem, fewer have tried to correct its impact. Many of these studies have looked at auditory environments and visual environments in isolation; even less studies have looked at audiovisual contexts. Thus, the fact that my method of audiovisual distance compression compensation is novel for various reasons:

- It is *more widely applicable* in the context of modern VR design. VR is moving away from traditional visual only displays and incorporating 3D audio,

haptics, and even smell in order to render high fidelity virtual environments.

- It is *implementation detail agnostic*: the method is theoretically independent of a given hardware configuration and software rendering system. Instead, it is grounded in the theory of multi-sensory perception, is psychophysically validated, and is simple to implement.

- It is *extensible*: While I focused on egocentric distance compression, incongruence is a technique that could be applied to the visual periphery. Experiment II alludes to a parabolic model of distance compression in VR with respect to the angle subtended from the observer's mid-sagittal plane to the object in question. While this requires further investigation, it is easy to see how the incongruence method could be easily extended by modelling this angle as a factor into the method.

The next few sections draw the results from Experiments I, II, and III in light of the research questions specified in Chapter 1.

RQ1: Can distance compression be compensated for in audiovisual virtual environments using incongruence?

The first research question posed in this thesis is whether distance compression can be compensated for. More specifically, can it be compensated for in a sense that its effects are minimized, with the ultimate goal being complete elimination of compression. The first study addressed this question in detail. Taking a novel approach, the first study grounds itself in the theory of maximum likelihood multi-sensory integration. Applying a controlled study allowed to get a low level interpretation of the integration scheme between audition and vision when participants were making their distance estimates. The results demonstrated a statistically significant effect of reducing the compression rate in participants as measured from the psychometric curves. When presented an incongruent environment, participants were more likely to choose a correct response in the 2AFC task than a congruent environment.

The method works by anchoring the audio components' position of an audiovisual object to the visual component, and then computing the offset of the audio

component. When the visual object is close to the user, around 8 meters, the incongruent function computes an offset behind the visual component. When it is closer than 8 meters, it brings the audio closer to the observer. This incongruent positioning was most effective in objects that were placed in the region of 5 to 10 meters from the observer. This indicates that the function should be modified for objects that are to be rendered closer to the observer. For small room simulations, it would be interesting to see how the function may be adapted to eliminate the perceived compression.

RQ2: Does the compensation function generalize to less abstract environments?

After establishing the incongruent method as a viable approach to reducing distance compression, the next research question was synthesized based on limitations and drawbacks from the environment used to test the RQ1. Namely, the environment was abstract, consisting of a green cube with some pink noise as an audiovisual object, rendered on a white plane in a blue sky box. The green cube and pink noise had no *semantic* relationship: they had nothing natural connecting them together. Also, the environment had minimal distance cues with respect to both vision and audition: the environment was empty with nothing but the experimental stimuli. These issues make it difficult to generalize the findings from the first study, and state that incongruence is an effective method for reducing distance compression in real use cases. This is a critical concern, as one of the key application domains of the method is in real time simulation of environments through virtual reality. It is therefore imperative to consider, test, and validate the method in more ecological environments. This need gave rise to the second experiment.

Thus, the second experiment was designed to incorporate some of the static cues from the literature base on distance compression. This was in order to simultaneously introduce richer distance cues while still maintaining a high degree of experimental control (i.e. the distance cues occlusion, contrast, shading, and relative size were all carefully controlled and included when choosing the photographic assets). The results gave greater insight into the application of incon-

gruence by identifying the need for a calibration phase to establish the degree of an individuals distance compression for a given virtual environment. While the compensation function showed a linear relationship with compression respect to egocentric distance, the function did not correct for the compression as the rate was too high. This implies that the rate should first be established for a given environment to set the parameters to the compensation function.

Another noteworthy aspect of Experiment II was the results for angular distance compression. While egocentric distance perception at the point of foveal vision, the visual periphery is also important to establish and parameterise in the compensation method. When mixed with audio, peripheral visual cues and out-of-view auditory cues can be used to notify or steer an observer's direction. Distance compression in peripheral vision is greatly understood, with few studies exploring this topic. Work from Mateeff & Gourevich and Bock looked at localization in the visual periphery, and their results showed systematic error in localization of stimuli presented in the periphery from the central gaze fixation [213, 214]. Given the phenomenon of distance compression, and the interaction between angle and distance observed in the second experiment, it is reasonable to suggest that such peripheral issues would be exasperated in VR. This is before multi-sensory perception with respect to audio is even considered. It is intriguing to ponder how these interactions may play out, and specifically design experiments to observe and track the interactions.

RQ3: Does the compensation function generalize to dynamic environments?

Having established how incongruence interacts with various distance cues in static environments, I then became interested in stepping up into the world of dynamic environments. Here, the question was how might incongruence interact with other sources of distance cues? The experiment specifically investigated two sources of cues. First, ones that come from our own sense of movement, termed proprioception. As we physically move through a space, familiarity of our own stride, muscle movement, and velocity all contribute to our sense of distance travelled. This information has been integrated and adapted to in the real world. Previous

studies have explored distance perception in the context of proprioception and observed similar compression as reported in this thesis [191, 203, 215].

Second, cues that come from the interaction between sound sources and the environment; reverberant distance cues where artificial echoes were applied to the sound stimulus. As the participant moved forward, either by walking or through the use of the gamepad, the artificial spatial environment updated with respect to not only the graphical display but the intensity of the sound and the intensity of the echoes. Reverberation has been shown to be an absolute distance cue [9, 131]. The expectation was that with reverberation acting as an absolute cue to distance, combined with the intensity change as the participant walked forward, would result in a better distance discrimination and reduced stopping distance.

The results of the study were mixed. While reverberation was expected to impact participants' judgement of the distance to the stimuli, there was no statistically significant effect observed. It is unclear why this occurred, as the literature on distance perception describes reverberation as an absolute cue (See work from Mershon & King and Zahorik for evidence of loudness constancy [131, 7], and Altmann et al. for insight at a neuronal level [216]). One hypothesis is the simplicity of the reverberation filter applied to the signal. Simple Schroeder reverberation was implemented [210]. This reverberation is artificial, and it operates by simply applying an all pass filter combined with a delay circuit. While perceptually giving a sense of depth to the signal, it does not match with the visual representation of the virtual environment rendered. As the participant walked closer to the sound source, one would expect the echoes to change. However, in this simple model that doesn't happen naturally. The main result of incongruence as an effective compensator for distance compression was not observed. The incongruity function is derived from studies that have aimed at reproducing the captured impulse responses [139] in static environments. Experiment III employed a dynamic environment and suggests a more complex incongruence function is required to accommodate in dynamic environments.

7.3 Limitations & Future Work

All three studies combined have measured the impact of incongruence from static, abstract environments on to cue rich scenes, finally to simplistic dynamic environments. It is important however to discuss the limitations of these findings. Below I list the contextual constraints that give way to limitations of the findings, and exemplify future work in the area of distance compression compensation.

To begin with, while both static and dynamic environments have been explored, they still had characteristics in common. Namely, in all studies there existed a single observer and, with the exception of study two, a single target in the centre field of view. Thus it is difficult to qualify the efficacy of incongruence when the environment is more populated. Very little research has been made into distance perception in densely populated virtual environments. Some researchers have studied distance perception in cross-modal real world environments with multiple targets [138], while others have used perceptual matching techniques within and across modalities [120, 130]. Others have applied a similar technique to the one in Experiment I, that of psychophysical measurements and analysis [89, 121]. The perceptual matching technique has been shown to be an effective calibration procedure, and involves placing more than one object in the field of view, with the observers task typically being to align both objects, yet are typically applied to augmented reality environments. Analogous methods in virtual environments can be the method of adjustment or forced choice experiments.

Experiment II began the discussion on the interaction between incongruence and localization. A statistically significant interaction was observed between the distance adjustments made by participants in the study and the azimuth angle of the particular target. This led to the hypothesis that a more generally applicable compression compensation function is needed. Such a function would need to factor in the target location angle from the observer's centre point of view. Also, as the ventriloquist effect has been shown to reverse in heavily blurred environments (with audition 'capturing' vision) [103], it is worthwhile to consider how the degree of incongruence should be altered. A psychophysical study would unveil the weighting schemes applied in this scenario, and lead to a deeper understanding of the integration process resulting in the three dimensional localized percept. As

Experiment III showed, in mobile VR environments, other sources of vision such as proprioception would need to be factored in also. Future work could explore this in detail through a rigorous experimental design.

Another limitation is the accuracy to which these environments presented their distance cues. With respect to visual cues, all environments were rendered using the default rendering algorithms provided by Oculus Technologies as part of the Unity Integration SDK[1]. In VR systems, there are always other sources of error such as the display field of view, the rendering and update latency, and the resolution of the head mounted display. All were controlled for by using the same hardware in all three experiments, yet as display type has been shown to interact with distance perception [163, 217, 218], it is imperative to test incongruence using various display formats, not just various HMDs. With respect to audible cues, intensity and reverberation were employed in the experiments, however these were not rendered by simulating real world dynamics. Intensity drop and gain followed a $\frac{1}{d}$ law where d is the distance from the observer. This is the physical relationship observed with respect to pressure in the real world, however the reverberation applied was artificial. Signal processing methods specifically in the domain of audio have been developed which capture a given acoustic field for analysis and reproduction [58, 219, 220]. Combining this with results from perceptual matching in VR is exciting, as it will lead to developments in applying incongruence to more natural simulations of the real world. This is necessary for commercial VR applications that aim to recreate scenarios such as the office environment, virtual co-located presence etc.

To conclude, incongruence has been shown to interact with distance perception in audiovisual virtual environments. The technique is simple, based on an empirically tested theory of multi-sensory integration. In my work, I have studied incongruence with respect to egocentric distance perception, by designing three studies manipulating aspects of the virtual environment. These studies have resulted in novel findings regarding both human spatial perception within audiovisual virtual environments, as well as highlighting implications for VE designers. All software was built using custom code with open source components, which have in turn been open sourced back to the wider community. I have discussed

[1]See Appendix A.

future work that could explore the topic further in the context of allocentric distance perception, more densely populated environments, and real scenario spatial cue capture.

7.4 Impact

The findings from this thesis have implications for applications and the design of VEs. VR is an exciting but immature field, and we are still learning the techniques required to implement successful VEs. Designers should create VEs that are carefully tailored to human spatial perception. Investigating incongruence to understand its potential effectiveness in compensating for distance compression can inform engineers in developing tools for VE designers to enhance the mapping between designer intentions and the user's perception. There is much work remaining in distance perception; perhaps one day we will identify a complete set of factors involved and derive a universal solution to the distance compression problem.

Irrespective of distance perception, there are implications for audio research on future VR applications. As we understand how humans perceive sounds in the real world, how we fuse together auditory information in order to make sense of our environment, this will feed directly into the hardware and software of VR. One clear impact is seen in studying the physics of sound propagation and human perception in tandem. To physically simulate sound dispersal in 3D environments in real time is computationally intractable[2] However, studies in human perception demonstrate that there is much in a soundscape that humans cannot even perceive. Therefore, if we apply human perception to the design of algorithms for 3D sound, we can greatly reduce the complexity which can result in a real time, realistic impression of 3D virtual worlds that suspend our disbelief. It is intriguing to think of all the exciting insights that will be found in the mean time as interdisciplinary work in this area continues, hopefully for decades to come.

[2]at the time of writing at least.

Appendices

Appendix A

Software Engineering Portfolio

An EngD differs from a PhD in many subtle ways, but the most obvious way is the dynamic of operating in a corporate versus an academic environment. Over the course of my EngD, I have written countless lines of code in various programming languages, contributing to many different projects. This appendix serves as a minor portfolio of all the projects I have worked on, from commercial enterprises to academic software development. My aim is not to list every single detail: instead, in this appendix, I aim to highlight the most important pieces of work I have developed, some of which have been highlighted in the main passage text of the thesis. These key pieces of software were used in conducting the experiments I've reported.

Thus in this appendix, I wish to disclose the details and inner workings of the core software I have developed. I have decided to include such detail in an appendix so as not to subtract from the research work I have presented. While ultimately I do believe software is a means to an end goal application, in order to then conduct some human centred studies, it is important not to neglect the journey that was taken. Technophiles may rejoice, while others can skip on by. The point is that the choice is yours.

A.1 Unity Audio Plug-in

When I began designing the experiment for my first study, it quickly became apparent that there was no readily usable solutions for spatial audio rendering inside virtual reality systems. A few companies, such as Visisonics had exhibited their RealSpace3D© Audio solution, yet this and many others remained closed source. This meant it was not feasible to use it in order to create the virtual environments I needed to conduct my experiments. In the meantime, there was a very active community of audio engineers in the research domain writing code to render spatial audio scenes for perceptual studies and engineering.

Unity, the game engine from Unity Technologies (from here on in referred to simply as Unity) is a very powerful and flexible engine, primarily targeted at the games industry for developing cross platform titles. The engine itself handles much of the low level platform specific functionality for rendering to the display, peripheral input, and file I/O. It encapsulates much of this functionality, doing the grunt work for the application developer, enabling them to focus on the core specifics of their game. The engine acts as a level designer, scripting engine, animator, renderer, compiler, and deployment module, meaning it streamlines much of the game development pipeline into a single, cross-platform utility. However, the engine remains highly customizable through a plug-in style system architecture: application developers can write their own platform specific, low level code if they so wish, effectively bypassing certain features of the engine. For example, application developers can implement their own rendering algorithm, or write a sub-module in C/C++ , compile it to a shared library, and make an interop call to the library from Unity at runtime.

Using this plug-in system, I built a complete, low level binaural engine and integrated it into Unity's runtime environment. Using the Rift SDK integration module from Oculus VR, I was then able to create a complete audiovisual VR system, which I then used to create the environments for all three of my studies. The code is publicly available, and is included in binary and source code form in the accompanying material to this thesis.

A.1.1 Engine Module Architecture

The Unity Audio Plug-in has a stacked system architecture; it consists of two main layers. The bottom layer is composed of 3 main parts; a soundbank, a binaural renderer, and a digital-to-analog decoder and I/O stream interface. The top layer is a lightweight interface layer to Unity itself. It consists of 3 C# scripts which handle the function calls to the low level C++ code, wrapping the functionality into an Object-Oriented design, which fits with Unity's object based programming scheme.

The plug-in itself is packaged as an OSX bundle[1] containing the C++ executable compiled against OSX 10.11. The bundle contains all the dependencies as prebuilt shared libraries (discussed in Section A.1.4), and can thus be compiled once for the platform, then just installed on any machine without the need for recompilation or installation of any other system libraries. As such, the plug-in also requires no working knowledge of C++ whatsoever. All interfacing is handled through the C# scripts in the top layer, meaning any competent C# programmer can pick it up and use it in their Unity application.

Figure A-1 shows the system architecture in a more digestible graphical format. The top layer is relatively trivial: it consists of a main script which implements the Unity $\leftrightarrow$ C++ interface as a MonoBehaviour script[2], and two other scripts, one used to instantiate 3D audio components of objects, and another which implements the concept of a listener. From the application developer's point of view, this layer is the one they interact with. As part of the plug-in, I implemented a test scene which shows how the C# scripts can be used. Once the

[1] As documented in the source code repository's README file at the time of publication, the plug-in is only available on OSX.

[2] This is a Unity specific concept; all interactive objects in Unity are modelled after entities in the environment, termed as scenes, that can have any number of behaviours attached to them as scripts. The objects appearance is modelled with a 3D mesh, its physical properties are passed into the physics engine in Unity, while the application specific custom behaviours are scripted by the application developer using the MonoBehaviour interface. 3D properties such as the model transformation as well as physical properties such as velocity and acceleration can be queried and modified through such behaviour scripts. The Unity Audio Plug-in uses a MonoBehaviour to wrap the low level functionality into a set of accessor methods. By attaching this behaviour script, an object in the Unity scene can then act as a 'Manager' to the audio renderer. Details about the MonoBehaviour interface, Objects, and Scenes can be found Unity's documentation and user manual.

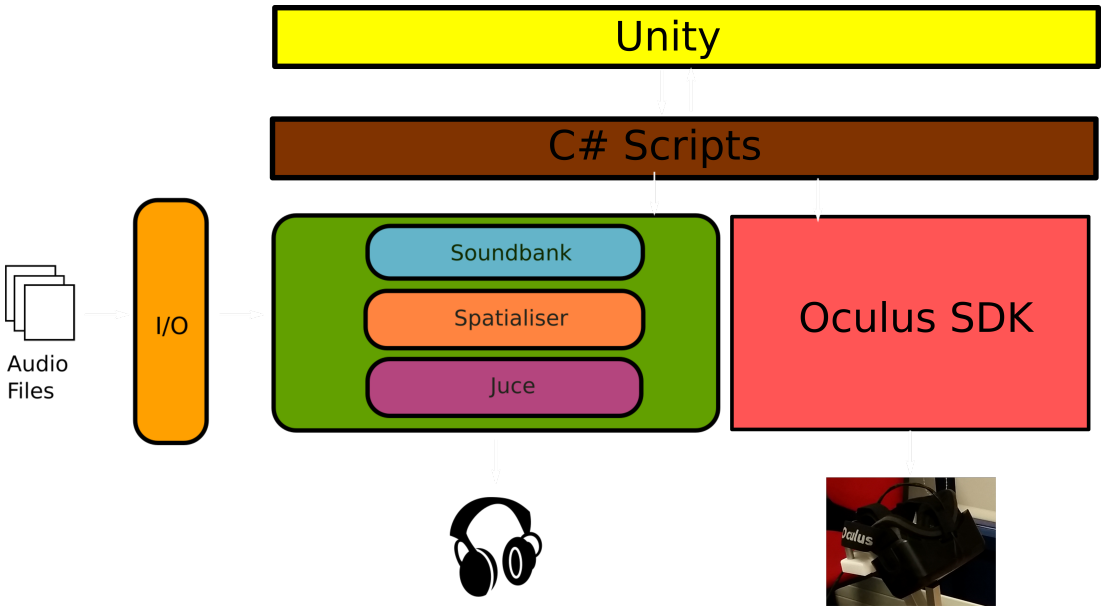

Figure A-1: Architecture for the plug-in showing the dual layer implementation, with the Oculus Rift SDK to the side. The bottom level interacts with the headphones and the top level's scripts in order to drive the binaural audio renderer.

plug-in has been imported, all that is required is to attach the AudioManager MonoBehaviour script to any object in the scene, then create a few AudioObject instances and an AudioListener in order to hear the audio spatialised in the scene.

A.1.2 Binaural Rendering with the SSR

The Unity Audio Plug-in uses the SoundScape Renderer as its binaural renderer [38]. The SSR is a set of renderers for 3D audio. It implements Wave Field Synthesis, Ambisonics, and Binaural Rendering (See Chapter 2). In creating the Unity Audio Plug-in I used the SSR's binaural rendering capabilities. In this section I detail how to perform the convolution process required to process audio data and produce a binaural stereo mix.

The ssr::BinauralRenderer is a subclass of the apf::MimoProcessor, which is an abstract multiple input/multiple output processor that enables the programmer to implement the processing callback while handling the threading and access

control of the samples held in the processor's buffer. The renderer is instantiated by passing a apf::parameter_map instance which is a key-value dictionary of configuration settings for the renderer. The main settings required are the sample rate, block size and the location (full path) to the HRIR file that contains the impulse responses for convolution.

The processor functions through use of the Policy Design Pattern. This design pattern dictates that a can have a number of different policies for responding to similar situations (i.e. a class may have a number of different policies regarding the printing of data to a file or to a TCP stream). In my case, the policies that the renderer is concerned with are its interface (how to process it's data buffers at each audio cycle or rather how to 'use' and interface with it) and how to act in a threaded manner. To specify which policies to use, you simply include the header file of the policy and define a macro called `APF_MIMOPROCESSOR_INTERFACE_POLICY` for the interface policy and `APF_MIMOPROCESSOR_THREAD_POLICY` for the thread policy. The library comes with a default thread header which just uses a single threaded policy for both on windows (currently not implemented) and the POSIX library for Unix and OSX systems.

In order to use the binaural renderer, you need to specify the policies you want to use. The renderer relies on two policies in it's implementation; an interface policy and a threading policy. To use the renderer as a standalone module, the pointer policy must be used. This then opens up the `audioCallback` function to be called manually by the application programmer when they want to process some data. The function accepts 3 arguments, the block size of the frame to be processed, a pointer to a series of inputs and a pointer to a pair of outputs (as the binaural renderer is an instance of an N-input, 2-output processor for stereo binaural output).

The processor requires it's input to be a pointer to a list of channels. These channels can be implemented as a series of vectors. The renderer's output is also expected to be a series of vectors representing the audio channels. The inputs should be an `N * BLOCK_SIZE` matrix where `BLOCK_SIZE` is the number of frames to be processed as a block during each run of the audio cycle. The N is

the number of channels in the input. The outputs should be a `2 * BLOCK_SIZE` matrix, indicating stereo output. The renderer expects a 1-1 mapping of input channels to sources in order (i.e. `channels[0]` is the first source, `channels[1]` the second etc...).

While this process operates in real time, it can be used for encoding a set of mono sources into a binaural stereo file. In order to create a binaural file, the channels need to be transposed as libsndfile reads in row-wise order, interleaving the channels as it dumps them to the file.

In summary:

1. The binaural renderer can be instantiated after specifying the policies required
2. Next, one must generate a parameter map, a key-value dictionary containing the configuration (block size, HRIR file path etc.) for the renderer.
3. Pass a pointer to a list of arrays representing the channels of the audio (best to use the `apf::fixed_matrix` container that comes with the APF framework).
4. If dumping to a file, a second output buffer, which is the transposed matrix of the output list of channels, is required as libsndfile expects reads and writes in row wise order for (de)interleaving. One can then call the `writef()` function of the `SndFileHandle` object in the libsndfile C++ API to write the stereo output to a file.

A.1.3 Digital-to-Analog Conversion & Soundbank Functionality

In order to handle the output from the renderer to the headphones, I used the digital-to-analog converter (DAC) module from the JUCE library[3]. The `jucePlayer.h` header file contains a declaration of a C++ class that hooks into Juce's Audio IO callback system. Juce handles the threading of the audio player

[3] `https://www.juce.com/`

itself internally; all that is required is to extend the abstract class `AudioIODeviceCallback` and implement the virtual `audioDeviceIOCallback()` method. The `AudioIODeviceCallback` class also contains more methods to interface with the IO engine in more detail. Application developers can hook callbacks to implement some application specific functionality when the IO device starts, is stopped, or to listen for and react to errors in the IO. As the Unity Audio Plug-in does not support microphone input for audio recording, most of these hooks are ignored, however it could be extended to do so, for example if an application developer wished to do some real world noise detection for in game use, or speech input.

The `SoundBank` module is a small set of classes that implement an in-memory bank of sounds. Sound sources are modelled as objects with a pointer to the next audio frame, an (X, Y) position, and a toggle to turn the audio on and off. The bank is a map of numerical identifiers to sound source instances. It enables indexing through via sound source name, as well as activating and deactivating sound sources. All manipulation of sound sources is made through the `SoundBank` instance. The class is wrapped in a thin layer of functions which define the interface between the C# scripts in the top layer. Functionality for adding sounds to the bank by name, activating and deactivating sounds, resetting sounds, and moving sounds is exposed through this layer. This thin layer instantiates the `AudioPlayer`, `SoundBank`, `Spatializer` classes as static objects at load time. These objects are stored in smart pointers to manage their life-cycle through an init/shut-down function pair.

A.1.4 Package Distribution

The Unity Audio Plug-in has many various external dependencies. These are distributed as pre-built shared libraries, except for the boost library. The boost is a library for cross-platform C++ development, containing many useful frameworks and extensions to the C++ standard. The libsndfile library is included for portable and opaque audio file IO (i.e. the underlying file format structure is irrelevant; libsndfile will decode and return the audio frame data in any given file format). The libfftw library is also included, required by the SSR to perform convolution.

The Unity Audio Plug-in is distributed in source code form, with a project file for the Xcode integrated development environment (IDE). The project can either be built from within Xcode, or can be compiled through the terminal on OSX. The short `build.sh` shell script can be used to call the `xcodebuild` command which will build the project from the terminal prompt. The Xcode project file has some post script actions that package the software into an OSX bundle. This bundle houses the shared library built by the project, and is required for integrating with Unity. Alternatively, it can be used as a standalone module for binaural rendering in other projects.

Included in the projects repository is a sample Unity project demonstrating how to use the plug-in. This project can be used as a kick-starter to a project involving binaural audio, or just as a reference to how to use the plug-in. I have also included the Fabian head model HRTF dataset from the SSR. The HRTF is not configurable at runtime. Using different HRTF models requires having a model in the format specified in the the original paper by Ahrens et al. then recompiling the Unity Audio Plug-in [38]. Finally, stereo image assets are also included in the repository

A.2 Motion Tracking Experiment Software

While the same plug-in code was used to conduct the second study, the third study involved two machines networked together. One machine was used to track the participants motion and render the binaural audio. The other machine was used to driver rendering the experimental scene using Unity.

This experiment required custom software to achieve the following objectives:

1. Track the physical location of the participant. This was achieved through the integration of a simple computer vision technique using an off-the-shelf implementation and some dedicated hardware

2. Send the positional co-ordinates from the machine tracking the participant to the other machine rendering the environment.

3. Map the co-ordinates of the participant to the audio renderer in order to up-

date the incongruity function for positioning the audio and visual attributes of the audiovisual target in the environment.

A.2.1 Tracking the Participant

For tracking the participant, I used the excellent PSMove API (PSAPI) written by Thomas Perl[4]. The library implements a sensor fusion algorithm for tracking the position and orientation of a PSMove controller in 3D space. The Sony PSMove© controller (PSMove) was initially released as a peripheral for the PlayStation© 3 (PS3) system, allowing motion control. It consists of an accelerometer, a luminous LED orb that can light up in different colours, as well as buttons for interfacing with the PS3 system.

The PSAPI library incorporates a bluetooth API for talking to the controller, as well as libusb for reading images from the PSEye camera (PSEye), although any standard web camera is sufficient. The PSMove controller was attached to the top of the Oculus Rift, allowing one dimension of motion tracking, namely along the X-axis perpendicular to the camera's FOV. The library uses libusb to read image frames from the USB web camera, then the OpenCV[5] library for image processing. A high level platform independent Bluetooth protocol is used to query sensor data from the PSMove, reading orientation data from the accelerometer inside the controller, as well as toggling the state of the LED and querying the battery level. The PSMove is also equipped with a trigger button, and two front facing thumb buttons that can be used to select or interact with objects. These were not used for the studies in this thesis.

The PSAPI tracks the position of the PSMove's bulb as it travels through a 3D volume bonded by the PSEye's FOV and the internal software parameters related to the radius of the bulb. X and Y positions are tracked by searching through each frame from the PSEye for a circle of known colour matching the LED of the PSMove. The Z coordinate is determined by the current radius of the bulb in each frame. As the experimental design involved walking in 1 dimension perpendicular to the PSEye's FOV, this was not used. Only the X coordinate was used, as the

[4] `http://thp.io/2010/psmove/`
[5] `http://opencv.org/`

PSEye was positioned perpendicular to the direction participants were tasked to walk. Thus, by mounting the PSMove securely on to the top of the Oculus Rift (OR), as participants walked in perpendicular to the PSEye, their current position in space was tracked.

Once the coordinates of the person are known, these are then relayed to the machine in charge of rendering the virtual environment. In the experiment, both machines were connected together using an Ethernet cable, and the software implemented a Client-Server architecture. The server machine was configured to a static IP address, which the client then used to establish a Transmission Control Protocol (TCP) connection. TCP was handled by the boost library, a cross-platform C++ library. As soon as either the client or the server application were terminated, the connection was automatically closed.

The software framework also includes functionality for mapping the XY coordinates of the PSMove into YZ coordinates in the virtual environment. This is achieved using a simple JSON library that stores the coordinates of the PSMove, which are transmitted over the network and performing a developer configurable mapping. JSON is a human readable data format making it cross platform. It is widely used in networking applications. Its application in the framework I wrote facilitates complete decoupling of the tracking module from the OR rendering. This allows for future applications to plug in to different rendering frameworks, or for the same rendering software to use a different tracking system.

Appendix B

Thesis Examples Source Code

Many of the examples in early chapters of the thesis use graphs and data that are automatically generated by R scripts I have written. These scripts have been included in the thesis supplementary material. All instances of examples that use dummy data will reference this appendix meaning that source code lies in the repository.

Bibliography

[1] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth." *Handbook of perception and cognition, Vol 5: Perception of space and motion*, vol. 5, pp. 69–117, 1995.

[2] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert, "The perception of egocentric distances in virtual environments - A review," *ACM Computing Surveys*, vol. 46, no. 2, pp. 1–40, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2543581.2543590

[3] "Virtual reality head mounted displays (HMD) unit sales worldwide from 2014 to 2018 (in million)," 2017. [Online]. Available: https://www.statista.com/statistics/426429/hmd-virtuial-reality-unit-sales-worldwide/

[4] G. C. Burdea and P. Coiffet, *Virtual reality technology*. John Wiley & Sons, 2003, vol. 1.

[5] D. L. Schacter, D. T. Gilbert, and D. M. Wegner, *Psychology*. Worth Publishers, 2009. [Online]. Available: https://books.google.co.uk/books?id=-9x8dngFRe0C

[6] J. M. Loomis, J. A. Da Silva, N. Fujita, and S. S. Fukusima, "Visual space perception and visually directed action." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 4, pp. 906–921, 1992.

[7] P. Zahorik and F. L. Wightman, "Loudness constancy with varying sound source distance." *Nature Neuroscience*, vol. 4, no. 1, pp. 78–83, 2001.

[8] B. J. Mohler, J. L. Campos, M. B. Weyel, and H. H. Bülthoff, "Gait parameters while walking in a head-mounted display virtual environment and the real world," *Proceedings of the 13th Eurographics Symposium on Virtual Environments*, pp. 85–88, 2007. [Online]. Available: http://diglib.eg.org/handle/10.2312/PE.VE2007Short.085-088

[9] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, 2015. [Online]. Available: http://link.springer.com/10.3758/s13414-015-1015-1

[10] S. A. Kuhl, W. B. Thompson, and S. H. Creem-Regehr, "Minification influences spatial judgments in virtual environments," *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization - APGV '06*, vol. 1, no. 212, p. 15, 2006. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1140491.1140494

[11] H. Y. Kim, Y. Suzuki, S. Takane, and T. Sone, "Control of auditory distance perception based on the auditory parallax model," *Applied Acoustics*, vol. 62, no. 3, pp. 245–270, 2001.

[12] B. Carty, "Movements in Binaural Space : Issues in HRTF Interpolation and Reverberation , with applications to Computer Music Volume 2 of 2," PhD, National University of Ireland Maynooth, 2010.

[13] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using Binaural and Spectral Cues for Azimuth and Elevation Localization," *2008 Ieee/Rsj International Conference on Robots and Intelligent Systems, Vols 1-3, Conference Proceedings*, no. January 2016, pp. 2185–2190, 2008.

[14] J. Profita, T. G. Bidder, J. M. Optiz, and J. F. Reynolds, "Perfect pitch," *American Journal of Medical Genetics*, vol. 29, no. 4, pp. 763–771, apr 1988. [Online]. Available: http://doi.wiley.com/10.1002/ajmg.1320290405

[15] American Standards Association, "Acoustical Terminology," American Standads Association, Tech. Rep., 1960. [Online]. Available: http:

//www.nssn.org/search/DetailResults.aspx?docid=338516{&}selnode=

[16] Á. Csapó and G. Wersényi, "Overview of auditory representations in human-machine interfaces," *ACM Computing Surveys*, vol. 46, no. 2, pp. 1–23, nov 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2543581.2543586

[17] D. K. McGookin, "Understanding and Improving the Identification of Concurrently Presented Earcons," PhD, University of Glasgow, 2004. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.110.2676

[18] M. Barron, "Taking account of loudness constancy for the loudness criterion for concert halls," *Applied Acoustics*, vol. 73, no. 11, pp. 1185–1189, nov 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0003682X12001624

[19] D. H. Mershon, D. H. Desaulniers, S. A. Kiefer, T. L. A. Jr, and J. T. Mills, "Perceived loudness and visually-determined auditory distance," *Perception*, vol. 10, no. 5, pp. 531–543, 1981. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7339572

[20] R. D. Melara and L. E. Marks, "Interaction among auditory dimensions: Timbre, pitch, and loudness," *Perception & Psychophysics*, vol. 48, no. 2, pp. 169–178, mar 1990. [Online]. Available: http://www.springerlink.com/index/10.3758/BF03207084

[21] W. Garner and G. L. Felfoldy, "Integrality of stimulus dimensions in various types of information processing," *Cognitive Psychology*, vol. 1, no. 3, pp. 225–241, aug 1970. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0010028570900162

[22] J. G. Neuhoff, J. Wayand, and G. Kramer, "Pitch and loudness interact in auditory displays: Can the data get lost in the map?" *Journal of Experimental Psychology: Applied*, vol. 8, no. 1, pp. 17–25, 2002. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.8.1.17

[23] S. Wilkie and T. Stockman, "The Perception of Auditory-Visual Looming in Film," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7900 LNCS, pp. 378–386. [Online]. Available: http://link.springer.com/10.1007/978-3-642-41248-6{_}21

[24] E. M. Wenzel, F. L. Wightman, and D. J. Kistler, "Localization with non-individualized virtual acoustic display cues," in *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*. New York, New York, USA: ACM Press, 1991, pp. 351–359. [Online]. Available: http://portal.acm.org/citation.cfm?doid=108844.108941

[25] E. M. Wenzel, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, p. 111, 1993. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/94/1/10.1121/1.407089

[26] M. Cohen and L. F. Ludwig, "Multidimensional audio window management," *International Journal of Man-Machine Studies*, vol. 34, no. 3, pp. 319–336, mar 1991. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/002073739190023Z

[27] M. Cohen, "Throwing, pitching and catching sound: audio windowing models and modes," *International Journal of Man-Machine Studies*, vol. 39, no. 2, pp. 269–304, aug 1993. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S002073738371062X

[28] N. Sawhney and C. Schmandt, "Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments," *ACM Transactions on Computer-Human Interaction*, vol. 7, no. 3, pp. 353–383, sep 2000. [Online]. Available: http://portal.acm.org/citation.cfm?doid=355324.355327

[29] G. Marentakis and S. A. Brewster, "A Study on Gestural Interaction with a 3D Audio Display," in *LNCS*, 2004, vol. 3160, pp. 180–

191. [Online]. Available: www.audioclouds.orghttp://link.springer.com/10.1007/978-3-540-28637-0{_}16

[30] R. Boonen, "An Offline Binaural Converting Algorithm For 3D Audio Contents: A Comparative Approach To The Implementation Using Channels and Objects," *AES 135th Convention*, 2013.

[31] R. Mehra, L. Antani, Sujeong Kim, and D. Manocha, "Source and Listener Directivity for Interactive Wave-Based Sound Propagation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 4, pp. 495–503, apr 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6777442

[32] S.-w. Jeon, Y.-c. Park, and D. H. Youn, "Auditory Distance Rendering Based on ICPD Control for Stereophonic 3D Audio System," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 529–533, may 2015. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6926755

[33] T. Funkhouser, N. Tsingos, and J. Jot, "Survey of methods for modeling sound propagation in interactive virtual environment systems," pp. 1–53, 2003. [Online]. Available: http://www.cs.princeton.edu/{~}funk/presence03.pdf

[34] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Audio Engineering Society*, vol. 144, pp. 357–360, 1997.

[35] J. Cofino, A. Barreto, and M. Adjouadi, "Comparing Two Methods of Sound Spatialization: Vector-Based Amplitude Panning (VBAP) Versus Linear Panning (LP)," in *Innovations and Advances in Computer, Information, Systems Sciences, and Engineering.* Springer New York, 2013, pp. 359–370. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-3535-8{_}31

[36] E. W. Start, "Direct Sound Enhancement by Wave Field Synthesis," p. 218, 1997. [Online]. Available: http://www.narcis.nl/publication/RecordID/oai:tudelft.nl:uuid:c80d5b58-67d3-4d84-9e73-390cd30bde0d

[37] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, p. 2764, 1993. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/93/5/10.1121/1.405852

[38] J. Ahrens, M. Geier, and S. Spors, "The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Audio Engineering Society Convention*. Audio Engineering Society, 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14460

[39] J. Müller, M. Geier, C. Dicke, and S. Spors, "The boomRoom," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. New York, New York, USA: ACM Press, 2014, pp. 247–256. [Online]. Available: http://dl.acm.org/citation.cfm?id=2556288.2557000

[40] R. Mehra, N. Raghuvanshi, L. Antani, A. Chandak, S. Curtis, and D. Manocha, "Wave-based sound propagation in large open scenes using an equivalent source formulation," *ACM Transactions on Graphics*, vol. 32, no. 2, pp. 1–13, apr 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2451236.2451245

[41] N. Raghuvanshi and J. Snyder, "Parametric wave field coding for precomputed sound propagation," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–11, jul 2014. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2601097.2601184

[42] J. Blauert, *The Technology of Binaural Listening*. Springer New York, 2013.

[43] L. Rayleigh, *Philosophical Magazine*. Taylor & Francis., 1907. [Online]. Available: https://books.google.co.uk/books?id=vVjKOdktZhsC

[44] F. L. Wightman and D. J. Kistler, "Headphone simulation of freefield listening. II: Psychophysical validation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 868–878, feb 1989.

[Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/85/2/10.1121/1.397558http://asa.scitation.org/doi/10.1121/1.397558

[45] ——, "Headphone simulation of freefield listening. I: Stimulus synthesis," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, feb 1989. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.397557

[46] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1997. [Online]. Available: https://mitpress.mit.edu/books/spatial-hearing

[47] N. I. Fisher, T. Lewis, and B. J. J. Embleton, *Statistical Analysis of Spherical Data*. Cambridge University Press, 1987. [Online]. Available: https://books.google.co.uk/books?id=yK4Quuq9tRgC

[48] R. Steingrimsson and R. D. Luce, "Evaluating a model of global psychophysical judgmentsII: Behavioral properties linking summations and productions," *Journal of Mathematical Psychology*, vol. 49, no. 4, pp. 308–319, aug 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0022249605000143

[49] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners." *Annual review of psychology*, vol. 42, pp. 135–159, 1991.

[50] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3-4, pp. 171–218, 1992. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0003682X9290046U

[51] J. M. Loomis, R. L. Klatzky, and R. G. Golledge, "Auditory Distance Perception in Real, Virtual, and Mixed Environments," *Mixed Reality : Merging Real And Virtual Worlds*, pp. 201–214, 1999.

[52] C. Mendonça, J. A. Santos, G. Campos, P. Dias, and J. P. Ferreira, "On the Impact of Training HRTF-Based Auralisation," *Interacção 2010 4ª Conferência Interacção Pessoa-Máquina*, pp. 1–5, 2010. [Online]. Available: http://webs.psi.uminho.pt/lvp/site/Publications{_}files/Mendonca{_}et{_}al{_}2010{_}Training{_}HRTF.pdf

[53] Y. Vazquez Alvarez and S. a. Brewster, "Designing spatial audio interfaces to support multiple audio streams," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services - MobileHCI '10*. New York, New York, USA: ACM Press, 2010, p. 253. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1851600.1851642

[54] W. G. Gardner, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, p. 3907, 1995. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/97/6/10.1121/1.412407

[55] M. J. Evans, "Analyzing head-related transfer function measurements using surface spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, p. 2400, oct 1998. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/104/4/10.1121/1.423749

[56] J. G. Richter, M. Pollow, F. Wefers, and J. Fels, "Spherical harmonics based hrtf datasets: Implementation and evaluation for real-time auralization," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 667–675, jul 2014. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article{&}issn=1610-1928{&}volume=100{&}issue=4{&}spage=667

[57] J. Chen, "A spatial feature extraction and regularization model for the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, p. 439, 1995. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/97/1/10.1121/1.413110

[58] C. Schissler, A. Nicholls, and R. Mehra, "Efficient HRTF-based Spatial Audio for Area and Volumetric Sources," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 4, pp. 1356–1366, apr 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7383327/

[59] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, p. 2821,

1999. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/105/5/10.1121/1.426898

[60] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, and H. Møller, "The Interaural Time Difference in Binaural Synthesis," in *Audio Engineering Society 108th Convention*, Paris, 2000, pp. 1–20. [Online]. Available: http://vbn.aau.dk/ws/files/227975878/2000{_}Minnaar{_}et{_}al{_}AES{_}Paris.pdf

[61] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*. IEEE, 2001, pp. 99–102. [Online]. Available: http://ieeexplore.ieee.org/document/969552/

[62] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Academic Press, 2000, no. April. [Online]. Available: https://human-factors.arc.nasa.gov/publications/Begault{_}2000{_}3d{_}Sound{_}Multimedia.pdf

[63] C. Cruz-Neira, D. J. Sandin, T. a. DeFanti, R. V. Kenyon, and J. C. Hart, "The CAVE: audio visual experience automatic virtual environment," *Communications of the ACM*, vol. 35, no. 6, pp. 64–72, 1992.

[64] J. M. Plumert, J. K. Kearney, J. F. Cremer, and K. Recker, "Distance perception in real and virtual environments," *ACM Transactions on Applied Perception*, vol. 2, no. 3, pp. 216–233, 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1077402

[65] D. Vogel and R. Balakrishnan, "Distant freehand pointing and clicking on very large, high resolution displays," in *Proceedings of the 18th annual ACM symposium on User interface software and technology - UIST '05*. New York, New York, USA: ACM Press, 2005, p. 33. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1095034.1095041

[66] I. V. Piryankova, S. De La Rosa, U. Kloos, H. H. Bülthoff, and B. J. Mohler, "Egocentric distance perception in large screen immersive

displays," *Displays*, vol. 34, no. 2, pp. 153–164, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.displa.2013.01.001

[67] T. Y. Grechkin, T. D. Nguyen, J. M. Plumert, J. F. Cremer, and J. K. Kearney, "How does presentation method and measurement protocol affect distance estimation in real and virtual environments?" *ACM Transactions on Applied Perception*, vol. 7, no. 4, pp. 1–18, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1823738.1823744

[68] A. L. Simeone, E. Velloso, and H. Gellersen, "Substitutional Reality," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. New York, New York, USA: ACM Press, 2015, pp. 3307–3316. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2702123.2702389

[69] M. McGill, D. Boland, R. Murray-Smith, and S. Brewster, "A Dose of Reality," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. New York, New York, USA: ACM Press, 2015, pp. 2143–2152. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2702123.2702382

[70] E. A. McManus, B. Bodenheimer, S. Streuber, S. De La Rosa, H. Bülthoff, and B. J. Mohler, "The influence of avatar (self and character) animations on distance estimation, object interaction and locomotion in immersive virtual environments," *Symposiumon Applied Perception in Graphics and Visualization - APGV 2011*, vol. 1, no. 212, pp. 37–44, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2077458

[71] M. Leyrer, S. A. Linkenauger, H. H. Bülthoff, U. Kloos, and B. Mohler, "The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments," in *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization - APGV '11*, vol. 1, no. 212. New York, New York, USA: ACM Press, 2011, p. 67. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2077451.2077464

[72] Q. Lin, J. Rieser, and B. Bodenheimer, "Affordance Judgments in HMD-Based Virtual Environments," *ACM Transactions on Applied*

Perception, vol. 12, no. 2, pp. 1–21, apr 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2746686.2720020

[73] B. J. Mohler, S. H. Creem-Regehr, W. B. Thompson, and H. H. Bülthoff, "The Effect of Viewing a Self-Avatar on Distance Judgments in an HMD-Based Virtual Environment," *Presence: Teleoperators and Virtual Environments*, vol. 19, no. 3, pp. 230–242, jun 2010. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/pres.19.3.230

[74] M. Obrist, C. Velasco, C. T. Vi, N. Ranasinghe, A. Israr, A. D. Cheok, C. Spence, and P. Gopalakrishnakone, "Touch, Taste, & Smell User Interfaces," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. New York, New York, USA: ACM Press, 2016, pp. 3285–3292. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2851581.2856462

[75] N. Ranasinghe, K.-Y. Lee, G. Suthokumar, and E. Y.-L. Do, "Taste+," in *Proceedings of the ACM International Conference on Multimedia - MM '14*. New York, New York, USA: ACM Press, 2014, pp. 737–738. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2647868.2654878

[76] V. Harrar and C. Spence, "The taste of cutlery: how the taste of food is affected by the weight, size, shape, and colour of the cutlery used to eat it," *Flavour*, vol. 2, no. 1, p. 21, 2013. [Online]. Available: http://www.flavourjournal.com/content/2/1/21/abstract

[77] M. Murer, I. Aslan, and M. Tscheligi, "LOLL io," in *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction - TEI '13*. New York, New York, USA: ACM Press, 2013, p. 299. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2460625.2460675

[78] M. Banzi, "Arduino Home Page," 2016. [Online]. Available: https://www.arduino.cc/

[79] T. Narumi, S. Nishizaka, T. Kajinami, T. Tanikawa, and M. Hirose, "Augmented reality flavors," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York,

New York, USA: ACM Press, 2011, p. 93. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1978942.1978957

[80] I. Iacovides, A. Cox, R. Kennedy, P. Cairns, and C. Jennett, "Removing the HUD: The Impact of Non-Diegetic Game Elements and Expertise on Player Involvement," in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*. New York, New York, USA: ACM Press, 2015, pp. 13–22. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2793107.2793120

[81] K. Knoblauch and L. T. Maloney, *Modeling Psychophysical Data in R*. New York, NY: Springer New York, 2012. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-4475-6

[82] C. C. Wier, W. Jesteadt, and D. M. Green, "A comparison of method-of-adjustment and forced-choice procedures in frequency discrimination," *Perception & Psychophysics*, vol. 19, no. 1, pp. 75–79, 1976.

[83] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory Distance Perception in Humans: A Summary of Past and Present Research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005. [Online]. Available: http://www.ingentaconnect.com/content/dav/aaua/2005/00000091/00000003/art00003

[84] E. Larsen and R. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. Wiley, 2004, vol. 1. [Online]. Available: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470858648.html

[85] F. M. Nieuwenhuizen, P. M. Zaal, M. Mulder, M. M. Van Paassen, and J. a. Mulder, "Modeling Human Multichannel Perception and Control Using Linear Time-Invariant Models," *Journal of Guidance, Control, and Dynamics*, vol. 31, no. 4, pp. 999–1013, jul 2008. [Online]. Available: http://arc.aiaa.org/doi/10.2514/1.32307

[86] C. W. Johnson, W. Dell, and C. Science, "Limitations of 3D Audio to Improve Auditory Cues in Aircraft Cockpits," in *International Systems*

Safety Conference, 2003. [Online]. Available: http://www.dcs.gla.ac.uk/{~}johnson/papers/ISSC2003/3daudio.pdf

[87] C. Spence, "Audiovisual multisensory integration," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 61–70, 2007.

[88] ——, "Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule," in *Annals of the New York Academy of Sciences*, vol. 1296, no. 1, aug 2013, pp. 31–49. [Online]. Available: http://doi.wiley.com/10.1111/nyas.12121

[89] J. P. Rolland, C. Meyer, K. Arthur, and E. Rinalducci, "Method of Adjustments versus Method of Constant Stimuli in the Quantification of Accuracy and Precision of Rendered Depth in Head-Mounted Displays," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 6, pp. 610–625, dec 2002. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/105474602321050730

[90] T. Kuroda and E. Hasuo, "The very first step to start psychophysical experiments," *Acoustical Science and Technology*, vol. 35, no. 1, pp. 1–9, 2014. [Online]. Available: http://jlc.jst.go.jp/DN/JST.JSTAGE/ast/35.1?lang=en{&}from=CrossRef{&}type=abstract

[91] A. Soranzo and M. Grassi, "Psychoacoustics: A comprehensive MATLAB toolbox for auditory testing," *Frontiers in Psychology*, vol. 5, no. JUL, pp. 1–13, 2014.

[92] D. M. Green, "A maximum-likelihood method for estimating thresholds in a yesno task," *The Journal of the Acoustical Society of America*, vol. 93, no. 4, p. 2096, 1993. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/93/4/10.1121/1.406696

[93] H. Dj, "Signal Detection Theory," *Copyright*, pp. 1–10, 1997.

[94] H. Wickham, "ggplot2 : Elegant graphics for data analysis," 2009. [Online]. Available: http://had.co.nz/ggplot2/book

[95] F. a. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit." *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[96] F. Steinicke, G. Bruder, J. Jerald, H. Frenz, and M. Lappe, "Estimation of detection thresholds for redirected walking techniques." *IEEE transactions on visualization and computer graphics*, vol. 16, no. 1, pp. 17–27, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19910658

[97] S. Spagnol, E. Tavazzi, and F. Avanzine, "Relative auditory distance discrimination with virtual nearby sound sources," in *18th International Conference on Digital Audio Effects (DAFx-15)*, Trondheim, 2015, pp. 1–6. [Online]. Available: http://www.soundofvision.net/relative-auditory-distance-discrimination-with-virtual-nearby-sound-sources/

[98] D. Senkowski, D. Talsma, M. Grigutsch, C. S. Herrmann, and M. G. Woldorff, "Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations," *Neuropsychologia*, vol. 45, no. 3, pp. 561–571, 2007.

[99] L. Piwek, F. Pollick, and K. Petrini, "Audiovisual integration of emotional signals from others' social interactions," *Frontiers in Psychology*, vol. 9, no. May, pp. 1–10, 2015. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00611

[100] D. W. Massaro, "A framework for evaluating multimodal integration by humans and a role for embodied conversational agents," in *Proceedings of the 6th international conference on Multimodal interfaces - ICMI '04*. New York, New York, USA: ACM Press, 2004, p. 24. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1027933.1027939

[101] D. Grelaud, N. Bonneel, M. Wimmer, M. Asselot, and G. Drettakis, "Efficient and practical audio-visual rendering for games using crossmodal perception," in *Proceedings of the 2009 symposium on Interactive 3D graphics and games - I3D '09*, vol. 1, no. 212. New York, New York, USA: ACM Press, 2009, p. 177. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1507149.1507178

[102] A. Engel, "Role of the temporal domain for response selection and perceptual binding," *Cerebral Cortex*, vol. 7, no. 6, pp. 571–582, sep 1997. [Online]. Available: https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/7.6.571

[103] D. Alais and D. Burr, "Ventriloquist Effect Results from Near-Optimal Bimodal Integration," *Current Biology*, vol. 14, no. 3, pp. 257–262, 2004.

[104] C. S. Choe, R. B. Welch, R. M. Gilford, and J. F. Juola, "The ventriloquist effect: Visual dominance or response bias?" *Perception & Psychophysics*, vol. 18, no. 1, pp. 55–60, jan 1975. [Online]. Available: http://www.springerlink.com/index/10.3758/BF03199367

[105] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, jun 1996. [Online]. Available: http://www.nature.com/doifinder/10.1038/381520a0

[106] H. MCGURK and J. MACDONALD, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, dec 1976. [Online]. Available: http://www.nature.com/doifinder/10.1038/264746a0

[107] V. Jousmäki and R. Hari, "Parchment-skin illusion: sound-biased touch." *Current biology : CB*, vol. 8, no. 6, p. R190, 1998.

[108] M. A. Meredith, J. W. Nemitz, and B. E. Stein, "Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors." *The Journal of neuroscience*, vol. 7, no. 10, pp. 3215–29, 1987. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3668625

[109] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002. [Online]. Available: http://www.nature.com/doifinder/10.1038/415429a

[110] M. A. Heller, "Haptic dominance in form perception with blurred vision," *Perception*, vol. 12, no. 5, pp. 607–613, 1983. [Online]. Available: http://pec.sagepub.com/lookup/doi/10.1068/p120607

[111] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, no. 6713, pp. 747–749, dec 1998. [Online]. Available: http://www.nature.com/doifinder/10.1038/25526

[112] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin, "Bayesian integration of visual and auditory signals for spatial localization." *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 20, no. 7, pp. 1391–1397, 2003.

[113] N. W. Roach, J. Heron, and P. V. McGraw, "Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration," *Proceedings of the Royal Society B: Biological Sciences*, vol. 273, no. 1598, pp. 2159–2168, sep 2006. [Online]. Available: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2006.3578

[114] B. Hervy, F. Laroche, J.-L. Kerouanton, A. Bernard, C. Courtin, L. D'haene, B. Guillet, and A. Waels, "Augmented historical scale model for museums," in *Proceedings of the 2014 Virtual Reality International Conference on - VRIC '14*. New York, New York, USA: ACM Press, 2014, pp. 1–4. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2617841.2617843

[115] R. S. Renner, E. Steindecker, M. Müller, B. M. Velichkovsky, R. Stelzer, S. Pannasch, and J. R. Helmert, "The Influence of the Stereo Base on Blind and Sighted Reaches in a Virtual Environment," *ACM Transactions on Applied Perception*, vol. 12, no. 2, pp. 1–18, mar 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2746686.2724716

[116] F. O. Matu, M. Thøgersen, B. Galsgaard, M. M. Jensen, and M. Kraus, "Stereoscopic augmented reality system for supervised training on minimal invasive surgery robots," in *Proceedings of the 2014 Virtual Reality International Conference on - VRIC '14*. New York, New York, USA: ACM Press, 2014, pp. 1–4. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2617841.2620722

[117] G. Parseihian, C. Jouffrais, and B. F. G. Katz, "Reaching nearby sources: comparison between real and virtual sound and visual targets," *Frontiers*

in Neuroscience, vol. 8, no. September, pp. 1–13, 2014.

[118] H. Wu, D. H. Ashmead, and B. Bodenheimer, "Using immersive virtual reality to evaluate pedestrian street crossing decisions at a roundabout," *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization - APGV '09*, vol. 1, no. 212, p. 35, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1620993.1621001

[119] S. H. Creem-Regehr, J. K. Stefanucci, and W. B. Thompson, "Perceiving Absolute Scale in Virtual Environments: How Theory and Application Have Mutually Informed the Role of Body-Based Perception," in *The Pschology of Learning and Motivation*, 2015, vol. 62, pp. 195–224. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0079742114000073

[120] J. E. Swan, M. A. Livingston, H. S. Smallman, D. Brown, Y. Baillot, J. L. Gabbard, and D. Hix, "A perceptual matching technique for depth judgments in optical, see-through augmented reality," *Proceedings - IEEE Virtual Reality*, vol. 2006, p. 3, 2006.

[121] J. P. Rolland, W. Gibson, and D. Ariely, "Towards Quantifying Depth and Size Perception in Virtual Environments," *Presence*, vol. 4, pp. 24–49, 1995. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/pres.1995.4.1.24{#}.V5d9fY7eM{_}Q

[122] A. Murgia and P. M. Sharkey, "Estimation of distances in virtual environments using size constancy," *The International Journal of Virtual Reality*, vol. 8, no. 1, pp. 67–74, 2009. [Online]. Available: http://centaur.reading.ac.uk/15341/

[123] R. Messing and F. H. Durgin, "Distance Perception and the Visual Horizon in Head-Mounted Displays," *ACM Transactions on Applied Perception*, vol. 2, no. 3, pp. 234–250, jul 2005. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1077399.1077403

[124] M. Paquier, N. Côté, F. Devillers, and V. Koehl, "Interaction between auditory and visual perceptions on distance estimations in a virtual environment," *Applied Acoustics*, vol. 105, pp. 186–199, apr

2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0003682X15003680

[125] J. F. Norman, O. C. Adkins, H. F. Norman, A. G. Cox, and C. E. Rogers, "Aging and the visual perception of exocentric distance," *Vision Research*, vol. 109, pp. 52–58, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698915000486

[126] T. Dat Nguyen, C. J. Ziemer, T. Grechkin, B. Chihak, J. M. Plumert, J. F. Cremer, and J. K. Kearney, "Effects of Scale Change on Distance Perception in Virtual Environments," *ACM Trans. Appl. Percept. Article*, vol. 8, no. 26, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2043603.2043608

[127] J. W. Kelly, W. Hammel, L. A. Sjolund, and Z. D. Siegel, "Frontal extents in virtual environments are not immune to underperception," *Attention, Perception, & Psychophysics*, vol. 77, no. 6, pp. 1848–1853, 2015. [Online]. Available: http://link.springer.com/10.3758/s13414-015-0948-8

[128] S. A. Kuhl, W. B. Thompson, and S. H. Creem-Regehr, "HMD calibration and its effects on distance judgments," *ACM Transactions on Applied Perception*, vol. 6, no. 3, pp. 1–20, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1577755.1577762

[129] R. E. Patterson, Ph.D., *Human Factors of Stereoscopic 3D Displays*. London: Springer London, 2015. [Online]. Available: http://link.springer.com/10.1007/978-1-4471-6651-1

[130] K. Ponto, M. Gleicher, R. G. Radwin, and H. J. Shin, "Perceptual calibration for immersive display environments." *IEEE transactions on visualization and computer graphics*, vol. 19, no. 4, pp. 691–700, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23428454

[131] D. H. Mershon and E. L. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975. [Online]. Avail-

able: http://link.springer.com/article/10.3758/BF03204113{%}5Cnhttp://www.springerlink.com/index/10.3758/BF03204113

[132] D. O. Kim, P. Zahorik, L. H. Carney, B. B. Bishop, and S. Kuwada, "Auditory Distance Coding in Rabbit Midbrain Neurons and Human Perception: Monaural Amplitude Modulation Depth as a Cue," *Journal of Neuroscience*, vol. 35, no. 13, pp. 5360–5372, apr 2015. [Online]. Available: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3798-14.2015

[133] A. Rungta, S. Rust, N. Morales, R. Klatzky, M. Lin, and D. Manocha, "Psychoacoustic Characterization of Propagation Effects in Virtual Environments," *ACM Transactions on Applied Perception*, vol. 13, no. 4, pp. 1–18, jul 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2974016.2947508

[134] P. Zahorik, "Assessing auditory distance perception using virtual acoustics." *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[135] S. Werner and J. Liebetrau, "Effects of shaping of binaural room impulse responses on localization," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, jul 2013, pp. 88–93. [Online]. Available: http://ieeexplore.ieee.org/document/6603216/

[136] M. S. Gordon, F. A. Russo, and E. MacDonald, "Spectral information for detection of acoustic time to arrival," *Attention, Perception, & Psychophysics*, vol. 75, no. 4, pp. 738–750, may 2013. [Online]. Available: http://link.springer.com/10.3758/s13414-013-0424-2

[137] J. M. Speigle and J. M. Loomis, "Auditory distance perception by translating observers," in *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*. IEEE Comput. Soc. Press, 1993, pp. 92–99. [Online]. Available: http://ieeexplore.ieee.org/document/378257/

[138] J. S. Chan, C. Maguinness, D. Lisiecka, A. Setti, and F. N. Newell, "Evidence for crossmodal interactions across depth on target localisation performance in a spatial array," *Perception*, vol. 41, no. 7, pp. 757–773, 2012.

[139] P. W. Anderson and P. Zahorik, "Auditory/visual distance estimation: accuracy and variability," *Frontiers in Psychology*, vol. 5, pp. 1–11, 2014. [Online]. Available: http://www.frontiersin.org/Auditory{_}Cognitive{_}Neuroscience/10.3389/fpsyg.2014.01097/abstract

[140] D. J. Finnegan, E. O'Neill, and M. Proulx, "Compensating for Distance Compression in Audiovisual Virtual Environments Using Incongruence," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 200–212, 2016.

[141] J. M. Loomis and J. W. Philbeck, "Measuring spatial perception with spatial updating and action," in *Embodiment, Ego-Space, and Action*, ser. Carnegie Mellon Symposia on Cognition, M. B. Roerta L. Klatzky Brian MacWhinney, Ed. Taylor and Francis, 2008, ch. 1, pp. 1–43. [Online]. Available: http://www.tandf.net/books/details/9780805862881/

[142] K. M. Rand, M. R. Tarampi, S. H. Creem-Regehr, and W. B. Thompson, "The Importance of a Visual Horizon for Distance Judgments under Severely Degraded Vision," *Perception*, vol. 40, no. 2, pp. 143–154, 2012.

[143] C. J. Lin, B. H. Woldegiorgis, D. Caesaron, and L.-Y. Cheng, "Distance estimation with mixed real and virtual targets in stereoscopic displays," *Displays*, vol. 36, pp. 41–48, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0141938214000912

[144] D. Waller and A. R. Richardson, "Correcting distance estimates by interacting with immersive virtual environments: effects of task and available sensory information." *Journal of Experimental Psychology: Applied*, vol. 14, no. 1, pp. 61–72, 2008.

[145] S. A. Kuhl, S. H. Creem-Regehr, and W. B. Thompson, "Individual differences in accuracy of blind walking to targets on the floor," *Journal of Vision*, vol. 6, no. 6, pp. 726–726, 2006. [Online]. Available: http://jov.arvojournals.org/Article.aspx?doi=10.1167/6.6.726

[146] J. M. Knapp and J. M. Loomis, "Visual Perception of Egocentric Distance in Real and Virtual Environments," in *Virtual and Adaptive*

Environments. CRC Press, jun 2003, pp. 21–46. [Online]. Available: http://dx.doi.org/10.1201/9781410608888.pt1

[147] C. S. Sahm, S. H. Creem-Regehr, W. B. Thompson, and P. Willemsen, "Throwing versus walking as indicators of distance perception in similar real and virtual environments," *ACM Transactions on Applied Perception*, vol. 2, no. 1, pp. 35–45, 2005.

[148] P. Willemsen, A. A. Gooch, W. B. Thompson, and S. H. Creem-Regehr, "Effects of Stereo Viewing Conditions on Distance Perception in Virtual Environments," *Presence: Teleoperators and Virtual Environments*, vol. 17, no. 1, pp. 91–101, 2008.

[149] S. H. Creem-Regehr, P. Willemsen, A. A. Gooch, and W. B. Thompson, "The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments," *Perception*, vol. 34, no. 2, pp. 191–204, 2005. [Online]. Available: http://pec.sagepub.com/lookup/doi/10.1068/p5144

[150] B. Wu, Z. J. He, and T. Leng, "Evidence for a sequential surface integration process hypothesis from judging egocentric distance with restricted view of the ground," *Journal of Vision*, vol. 3, no. 9, pp. 500–500, mar 2010. [Online]. Available: http://jov.arvojournals.org/Article.aspx?doi=10.1167/3.9.500

[151] B. Ries, V. Interrante, L. Anderson, and J. Lindquist, "Presence, rather than prior exposure, is the more strongly indicated factor in the accurate perception of egocentric distances in real world co-located immersive virtual environments," *Proceedings of the 3rd symposium on Applied perception in graphics and visualization - APGV '06*, 2006.

[152] R. T. Held, E. A. Cooper, J. F. O'Brien, and M. S. Banks, "Using blur to affect perceived distance and size," *ACM Transactions on Graphics*, vol. 29, no. 2, pp. 1–16, mar 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1731047.1731057

[153] M. Gorzel, D. Corrigan, G. Kearney, J. Squires, and F. Boland, "Distance

Perception in Virtual Audio-Visual Environments," *25th UK Conference of the Audio Engineering Society: Spatial Audio In Today's 3D World*, pp. 1–8, 2012.

[154] J. L. Campos, J. S. Butler, and H. H. Bülthoff, "Multisensory integration in the estimation of walked distances," *Experimental Brain Research*, vol. 218, no. 4, pp. 551–565, 2012.

[155] A. R. Richardson and D. Waller, "Interaction With an Immersive Virtual Environment Corrects Users' Distance Estimates," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 49, no. 3, pp. 507–517, 2007. [Online]. Available: http://hfs.sagepub.com/content/49/3/507.abstract

[156] J. W. Kelly, L. S. Donaldson, L. A. Sjolund, and J. B. Freiberg, "More than just perception-action recalibration: walking through a virtual environment causes rescaling of perceived space." *Attention, perception & psychophysics*, vol. 75, pp. 1473–85, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23839015

[157] P. Willemsen, M. B. Colton, S. H. Creem-Regehr, and W. B. Thompson, "The effects of head-mounted display mechanical properties and field of view on distance judgments in virtual environments," *ACM Transactions on Applied Perception*, vol. 6, no. 2, pp. 1–14, 2004.

[158] C. Firestone, "How "Paternalistic" Is Spatial Perception? Why Wearing a Heavy Backpack Doesn't - and Couldn't - Make Hills Look Steeper," *Perspectives on Psychological Science*, vol. 8, no. 4, pp. 455–473, jul 2013. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1745691613489835

[159] E. Klein, J. E. Swan, G. S. Schmidt, M. a. Livingston, and O. G. Staadt, "Measurement protocols for Medium-Field distance perception in Large-Screen immersive displays," *Proceedings - IEEE Virtual Reality*, pp. 107–113, 2009.

[160] J. K. Stefanucci, K. T. Gagnon, C. L. Tompkins, and K. E. Bullock,

"Plunging into the pool of death: Imagining a dangerous outcome influences distance perception," *Perception*, vol. 41, no. 1, pp. 1–11, 2012. [Online]. Available: http://pec.sagepub.com/lookup/doi/10.1068/p7131

[161] T. L. Ooi, B. Wu, and Z. J. He, "Distance determined by the angular declination below the horizon." *Nature*, vol. 414, no. 1991, pp. 197–200, 2001.

[162] A. Pasqualotto and M. J. Proulx, "The role of visual experience for the neural basis of spatial cognition," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 4, pp. 1179–1187, 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0149763412000176

[163] J. G. P. Corujeira and I. Oakley, "Stereoscopic Egocentric Distance Perception: The Impact of Eye Height and Display Devices," *SAP '13 Proceedings of the ACM Symposium on Applied Perception*, 2013.

[164] M. Leyrer, H. H. Ulthoff, B. J. Mohler, S. A. Linkenauger, and H. H. Ulthoff, "Eye Height Manipulations: A Possible Solution to Reduce Underestimation of Egocentric Distances in Head-Mounted Displays," *ACM Transactions on Applied Perception*, vol. 12, no. 1, 2015. [Online]. Available: http://dx.doi.org/10.1145/2699254

[165] P. D. Coleman, "An analysis of cues to auditory depth perception in free space." *Psychological bulletin*, vol. 60, no. 3, pp. 302–315, 1963.

[166] G. V. Békésy, "The Moon Illusion and similar auditory phenomena," *The American Journal of Psychology*, vol. 62, no. 4, pp. 540–552, 1949. [Online]. Available: http://www.jstor.org/stable/1418558

[167] D. H. Ashmead, D. Leroy, and R. D. Odom, "Perception of the relative distances of nearby sound sources," *Perception & Psychophysics*, vol. 47, no. 4, pp. 326–331, jul 1990. [Online]. Available: http://www.springerlink.com/index/10.3758/BF03210871

[168] A. W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms." *Nature*, vol. 397, no. 6719, pp. 517–520, 1999.

[169] R. A. Butler, E. T. Levy, and W. D. Neff, "Apparent distance of sounds recorded in echoic and anechoic chambers." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 6, no. 4, pp. 745–750, 1980. [Online]. Available: http://content.apa.org/journals/xhp/6/4/745

[170] D. S. Brungart, "Auditory localization of nearby sources. III. Stimulus effects," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, p. 3589, 1999. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/106/6/10.1121/1.428212

[171] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/112/5/10.1121/1.1506692

[172] D. H. Ashmead, D. L. Davis, and A. Northington, "Contribution of listeners' approaching motion to auditory distance perception." *Journal of experimental psychology. Human perception and performance*, vol. 21, no. 2, pp. 239–256, 1995. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-1523.21.2.239

[173] E. Hellier, J. Edworthy, B. Weedon, K. Walters, and A. Adams, "The Perceived Urgency of Speech Warnings: Semantics versus Acoustics," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 44, no. 1, pp. 1–17, jan 2002. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/0018720024494810

[174] S. Mateeff, J. Hohnsbein, and T. Noack, "Dynamic visual capture: apparent auditory motion induced by a moving visual target," *Perception*, vol. 14, no. 6, pp. 721–727, 1985. [Online]. Available: http://pec.sagepub.com/lookup/doi/10.1068/p140721

[175] E. H. Siegel and J. K. Stefanucci, "A little bit louder now: Negative affect increases perceived loudness." *Emotion*, vol. 11, no. 4, pp. 1006–1011, 2011. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0024590

[176] K. T. Gagnon, M. N. Geuss, and J. K. Stefanucci, "Fear influences

perceived reaching to targets in audition, but not vision," *Evolution and Human Behavior*, vol. 34, no. 1, pp. 49–54, jan 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1090513812000918

[177] M. Rychtáriková, T. van den Bogaert, G. Vermier, and J. Wouters, "Binaural sound source localization in real and virtual rooms," *Journal of the Audio Engineering Society*, vol. 57, no. 4, pp. 205–220, 2009. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14814

[178] G. Kearney, M. Gorzel, F. Boland, and H. Rice, "Depth Perception in Interactive Virtual Acoustic Environments Using Higher Order Ambisonic Soundfields," *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics, 2010, May 6-7, Paris, France*, 2010.

[179] D. Artaega, "An ambisonics decoder for irregular 3D loudspeaker arrays," in *Journal of the Audio Engineering Society*. Audio Engineering Society, 2013.

[180] D. H. Mershon, W. L. Ballenger, A. D. Little, P. L. McMurtry, and J. L. Buchanan, "Effects of room reflectance and background noise on perceived auditory distance," *Perception*, vol. 18, no. 3, pp. 403–416, 1989. [Online]. Available: http://pec.sagepub.com/lookup/doi/10.1068/p180403

[181] J. W. Philbeck and D. H. Mershon, "Knowledge about typical source output influences perceived auditory distance," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, p. 1980, 2002. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/111/5/10.1121/1.1471899

[182] M. G. Wisniewski, E. Mercado, K. Gramann, and S. Makeig, "Familiarity with speech affects cortical processing of auditory distance cues and increases acuity," *PLoS ONE*, vol. 7, no. 7, 2012.

[183] P. Jaekl, J. Seidlitz, L. R. Harris, and D. Tadin, "Audiovisual Delay as a Novel Cue to Visual Distance," *Plos One*, vol. 10, no. 10, p. e0141125, 2015. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0141125

[184] B. E. Stein and T. R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," *Nature Reviews*

Neuroscience, vol. 9, no. 4, pp. 255–266, 2008. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrn2331

[185] H.-J. Maempel and M. Jentsch, "Auditory and Visual Contribution to Egocentric Distance and Room Size Perception," *Building Acoustics*, vol. 20, no. 4, pp. 383–402, dec 2013. [Online]. Available: http://multi-science.atypon.com/doi/10.1260/1351-010X.20.4.383

[186] D. Kahneman, "Thinking Fast and Slow," p. 512, 2011.

[187] C. Sedikides, D. Ariely, and N. Olsen, "Contextual and Procedural Determinants of Partner Selection: Of Asymmetric Dominance and Prominence," *Social Cognition*, vol. 17, p. 118, 1999.

[188] S. Frederick, L. Lee, and E. Baskin, "The Limits of Attraction," *Journal of Marketing Research*, vol. 51, no. 4, pp. 487–507, aug 2014. [Online]. Available: http://journals.ama.org/doi/abs/10.1509/jmr.12.0061

[189] S. Füg, S. Werner, and K. Brandenburg, "Controlled Auditory Distance Perception using Binaural Headphone Reproduction Algorithms and Evaluation," 2012.

[190] Z. Zhou, A. D. Cheok, X. Yang, and Y. Qiu, "An experimental study on the role of software synthesized 3D sound in augmented reality environments," *Interacting with Computers*, vol. 16, no. 5, pp. 989–1016, 2004.

[191] H.-J. Sun, J. L. Campos, and G. S. W. Chan, "Multisensory integration in the estimation of relative path length," *Experimental Brain Research*, vol. 154, no. 2, pp. 246–254, 2004.

[192] A. Turner, J. Berry, and N. Holliman, "Can the perception of depth in stereoscopic images be influenced by 3D sound?" *Displays*, vol. 7863, no. February, 2011. [Online]. Available: http://link.aip.org/link/PSISDG/v7863/i1/p786307/s1{&}Agg=doi

[193] D. Bavelier, M. W. Dye, and P. C. Hauser, "Do deaf individuals see better?" *Trends in Cognitive Sciences*, vol. 10, no. 11, pp. 512–518, 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364661306002439

[194] M. A. García-Pérez, "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," *Vision Research*, vol. 38, no. 12, pp. 1861–1881, 1998. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0042698997003404

[195] "R: A language and environment for statistical computing," 2015. [Online]. Available: https://www.r-project.org/

[196] L. E. Cameron, J. C. Tai, and M. Carrasco, "Covert attention affects the psychometric function of contrast sensitivity," *Vision Research*, vol. 42, no. 8, pp. 949–967, 2002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0042698902000391

[197] G. Bruder, P. Lubas, and F. Steinicke, "Cognitive Resource Demands of Redirected Walking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 4, pp. 539–544, apr 2015. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7036075

[198] M. Grassi and A. Soranzo, "MLP: a MATLAB toolbox for rapid and reliable auditory threshold estimation." *Behavior research methods*, vol. 41, no. 1, pp. 20–28, 2009.

[199] G. Bruder, A. Pusch, and F. Steinicke, "Analyzing effects of geometric rendering parameters on size and distance estimation in on-axis stereographics," in *Proceedings of the ACM Symposium on Applied Perception - SAP '12*, vol. 1, no. 212. New York, New York, USA: ACM Press, 2012, p. 111. [Online]. Available: http://dl.acm.org/citation.cfm?id=2338699http://dl.acm.org/citation.cfm?doid=2338676.2338699

[200] N. A. Dodgson, "Variation and extrema of human interpupillary distance," in *International Society for Optics and Photonics XI*, A. J. Woods, J. O. Merritt, S. A. Benton, and M. T. Bolas, Eds., vol. 5291, no. July, may 2004, pp. 36–46. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=836662

[201] A. S. Gilinsky, "Perceived Size and Distance in Visual Space," *Psychological Review*, vol. 58, pp. 460 – 482, 1951.

[202] I. Frissen, J. L. Campos, J. L. Souman, and M. O. Ernst, "Integration of vestibular and proprioceptive signals for spatial updating," *Experimental Brain Research*, vol. 212, no. 2, pp. 163–176, jul 2011. [Online]. Available: http://link.springer.com/10.1007/s00221-011-2717-9

[203] H. J. Sun, J. L. Campos, M. Young, G. S. W. Chan, and C. G. Ellard, "The contributions of static visual cues, nonvisual cues, and optic flow in distance estimation," *Perception*, vol. 33, no. 1, pp. 49–65, 2004.

[204] S. Razzaque, Z. Kohn, and M. C. Whitton, "Redirected Walking," in *Eurographics 2001 - Short Presentations*. Eurographics Association, 2001, pp. 2–7. [Online]. Available: https://diglib.eg.org/handle/10.2312/egs20011036

[205] F. Steinicke, G. Bruder, L. Kohli, J. Jerald, and K. Hinrichs, "Taxonomy and Implementation of Redirection Techniques for Ubiquitous Passive Haptic Feedback," in *2008 International Conference on Cyberworlds*. IEEE, sep 2008, pp. 217–223. [Online]. Available: http://ieeexplore.ieee.org/document/4741303/

[206] R. L. Koslover, B. T. Gleeson, J. T. de Bever, and W. R. Provancher, "Mobile Navigation Using Haptic, Audio, and Visual Direction Cues with a Handheld Test Platform," *IEEE Transactions on Haptics*, vol. 5, no. 1, pp. 33–38, jan 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6060820/

[207] M. Gröhn, T. Lokki, and T. Takala, "Comparison of auditory, visual, and audiovisual navigation in a 3D space," *ACM Transactions on Applied Perception*, vol. 2, no. 4, pp. 564–570, oct 2005. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1101530.1101558

[208] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc*, vol. 49, no. 10, pp. 904–916, 2001. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10175

[209] D. Cabrera and D. Gilfillan, "Auditory distance perception of speech in the presence of noise," *Int. Conf. on Auditory Display, Kyoto*, pp. 1–9, 2002. [Online]. Available: http://icad.org/websiteV2.0/Conferences/ICAD2002/proceedings/08{_}DensilCabrera2.pdf

[210] M. Schroeder and B. Logan, ""Colorless" artificial reverberation," *IRE Transactions on Audio*, vol. AU-9, no. 6, pp. 209–214, nov 1961. [Online]. Available: http://ieeexplore.ieee.org/document/1166351/

[211] A. J. Kolarik, S. Cirstea, and S. Pardhan, "Evidence for enhanced discrimination of virtual auditory distance among blind listeners using level and direct-to-reverberant cues," *Experimental Brain Research*, vol. 224, no. 4, pp. 623–633, 2013.

[212] P. W. Anderson and P. Zahorik, "Auditory and visual distance estimation," *Proceedings of Meetings on Acoustics*, vol. 12, p. 050004, 2011. [Online]. Available: http://scitation.aip.org/content/asa/journal/poma/12/1/10.1121/1.3656353http://asa.scitation.org/doi/abs/10.1121/1.3656353

[213] S. Mateeff and A. Gourevich, "Peripheral vision and perceived visual direction," *Biological Cybernetics*, vol. 49, no. 2, pp. 111–118, dec 1983. [Online]. Available: http://link.springer.com/10.1007/BF00320391

[214] O. Bock, "Localization of objects in the peripheral visual field," *Behavioural Brain Research*, vol. 56, no. 1, pp. 77–84, jul 1993. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/016643289390023J

[215] M. Slater, M. Usoh, and A. Steed, "Taking steps: the influence of a walking technique on presence in virtual reality," *ACM Transactions on Computer-Human Interaction*, vol. 2, no. 3, pp. 201–219, sep 1995. [Online]. Available: http://portal.acm.org/citation.cfm?doid=210079.210084

[216] C. F. Altmann, K. Ono, A. Callan, M. Matsuhashi, T. Mima, and H. Fukuyama, "Environmental reverberation affects processing of sound intensity in right temporal cortex," *European Journal of Neuroscience*, vol. 38, no. 8, pp. 3210–3220, 2013.

[217] J. A. Jones, E. A. Suma, D. M. Krum, and M. Bolas, "Comparability of narrow and wide field-of-view head-mounted displays for medium-field distance judgments," in *Proceedings of the ACM Symposium on Applied Perception - SAP '12*. New York, New York, USA: ACM Press, 2012, p. 119. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2338676.2338701

[218] S. Boustila, A. Capobianco, and D. Bechmann, "Evaluation of factors affecting distance perception in architectural project review in immersive virtual environments," in *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology - VRST '15*. New York, New York, USA: ACM Press, 2015, pp. 207–216. [Online]. Available: http://dx.doi.org/10.1145/2821592.2821595http://dl.acm.org/citation.cfm?doid=2821592.2821595

[219] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16664

[220] J. Ahrens and S. Spors, "Reproduction of Moving Virtual Sound Sources with Special Attention to the Doppler Effect," *AES 124th Convention*, no. January, pp. 1–12, 2008.